Perception & Psychophysics
1995, 57 (2), 246–254

Notes and Comment

A model for realism of confidence judgments: Implications for underconfidence in sensory discrimination

WILLIAM R. FERRELL
University of Arizona, Tucson, Arizona

In a recent issue of this journal, Björkman, Juslin, and Winman (1993) presented a model of the calibration of subjective confidence judgments for sensory discrimination which they called "subjective distance theory." They proposed that there was a robust underconfidence bias in such judgments, that the model predicted such a bias, and that two different models were needed for the calibration of subjective confidence for cognitive judgments and for sensory ones. This paper addresses issues they raised. It points out that they have not presented a new model, but rather a portion of a more general one, the "decision-variable partition model" originally proposed in Ferrell and McGoey (1980). This paper explores properties of the model and shows, contrary to Björkman, Juslin, and Winman's hypotheses, that the model does not predict underconfidence, that the "hard–easy effect" can be observed with sensory discriminations, and that the model fits not only sensory, but also cognitive judgments.

In a recent issue of this journal, Björkman, Juslin, and Winman (1993) presented a model of the calibration of subjective confidence judgments for sensory discrimination, which they called "subjective distance theory." In that paper, they claimed that the model explains or supports their conclusions that there is a robust underconfidence bias in such judgments, that the "hard–easy effect" in calibration does not occur with sensory discriminations, and that different models are needed for the calibration of subjective confidence for cognitive judgments and for sensory ones. The model they presented is not new, and it does not support those conclusions. This paper seeks to clarify the properties of the model and to address the following issues that Björkman et al. have raised:

1. The model of calibration of subjective probability judgments they present is the same as the paired-comparison part of a more comprehensive model applying to both sensory and cognitive judgments in a variety of formats, of both full and limited range, proposed by me (Ferrell, 1994b; Ferrell & McGoey, 1978, 1980; Ferrell & Rehm, 1980; Smith & Ferrell, 1983). That model has been cited and described in reviews (Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1982; McClelland & Bolger, 1994). Originally termed "the decision-variable partition model," it is also known as "the signal-detection model" (McClelland & Bolger, 1994).

2. The model does not predict a general underconfidence bias. The model is quite consistent with the overconfidence often found in cognitive tasks and, indeed, was used by Ferrell and McGoey (1980) to explain that overconfidence as a failure of respondents to adjust their response criteria with changes in task difficulty.

3. The "hard–easy effect" has been observed with sensory judgments, and it is predicted by the model in appropriate circumstances.

4. The model fits calibration data for both sensory judgments and cognitive judgments, for example, for the subjective probability of answering general-knowledge questions correctly. It provides a framework within which the principal empirical observations about calibration in both domains can be explained. This suggests that both sensory judgments and some types of cognitive ones share a common process by which responses are assigned to the results of neural processing.

Review of the Decision-Variable Partition Model of Calibration of Subjective Probabilities

It is widely agreed that for consistency and practical applicability, numerical judgments of confidence in the form of subjective probabilities should be well calibrated, meaning that for all the events assigned a given probability, the proportion that actually occurs should equal the assigned probability. Calibration is clearly shown by a calibration curve, a graph of the proportion of occurrences as a function of the assigned probability, along with a graph of the distribution of responses. Statistics from scoring rules can be useful, but they obscure important features of the data.

Subjective probability questions can be asked in a variety of formats that any model of the process should attempt to take into account. These can often be characterized conveniently, though not uniquely, by the implied range of numerical responses and the nature of the event judged, the truth of a proposition or the correctness of a choice of alternatives, for example, "Is weight *a* heavier than weight *b*? Choose and give your subjective probability that you are correct." This is a half-range, probability-correct question. One should not give a probability less than .5 (otherwise, the other alternative should have been chosen), and the correctness of the choice is being judged. "Indicate which stimulus (of many possibilities) was presented and give your subjective probability that you are correct" could be either a

The author's mailing address is Systems and Industrial Engineering Department, University of Arizona, Tucson, AZ 85721 (e-mail: russ@sie.arizona.edu).

limited-range [1/(number of possibilities) to 1] probability-correct question or a full-range (0 to 1) probability-true question, depending on how the question is treated by the respondent.

The model for the calibration of answers to such questions that is presented in Ferrell and McGoey (1980) is based on signal-detection theory (Egan, Schulman, & Greenberg, 1959; Green & Swets, 1974; Swets,Tanner, & Birdsall, 1961). It assumes that there are two steps: (1) the generation of a magnitude of a scalar internal-decision variable that would enable one to decide the matter, that is, one that is monotone increasing with the probability of being correct, and (2) the association of that magnitude with a response. Different question formats require different decision variables.

The simplest case is that of the full-range, probability-true task, for example, "Give your subjective probability (on the range 0 to 1) that the stimulus was signal (rather than just noise)." The basic observation is of the single stimulus. That stimulus is assumed to produce a realization of an internal random observation variable having a greater mean value when it is signal than when it is noise. That internal variable resulting from the observation is, itself, the decision variable; the larger its magnitude, the more sure one can be that the stimulus event was a signal. The model assumes that there is a finite set of responses, that the range of the decision variable is partitioned into intervals by cutoff values, one interval for each response, and that responses are assigned to those intervals so that the next more extreme interval gets the next more extreme response. Responses are arbitrary and may be numerical probability values, verbal expressions, or actions. If the form and parameters of the distributions and the cutoffs are known or assumed, the calibration curve and the response proportions can be calculated as in Ferrell and McGoey (1980).

A somewhat more complicated case (and the one corresponding to the model presented in Björkman et al. (1993) is that of a two-alternative forced-choice task with the subjective probability response being the probability of correct choice, for example, "Which is heavier, weight *a* or *b*? Choose and give your subjective probability (from .5 to 1.0) that you are correct." Two observations are made, one for each alternative, producing two realizations of the internal variable, and the magnitude of the absolute difference between them is the decision variable that is partitioned to provide response categories. This case is considered in detail.

The basic observation is of a pair of stimuli (e.g., lifted weights). It generates two values of an internal (random) observation variable. The internal observation variable is assumed, without loss of generality, to be distributed with a higher mean for the larger (i.e., correct) alternative. In this case, the (optimal) decision variable is the *absolute value* of the difference between the two magnitudes of the observation variable; when that difference is large, one can have more confidence that the choice of the larger magnitude is correct. It must be the absolute value, because the respondent doesn't know which of the two alternatives is correct. (The distributions of the signed difference will depend upon the task, but it is frequently found that it may be assumed to be normal. That is a robust assumption, because the normal approximates other distributions and because the decision variable need only be determined within a monotone increasing transformation, so that if any such transformation can make the distribution normal, the assumption is satisfied.) Having partitioned the range of the magnitude of the decision variable, and having assigned the possible responses, in increasing order, to the successive intervals on it, the respondent chooses the alternative that gave the larger value of the observation variable. The respondent then considers the resulting magnitude of the decision variable, the absolute difference between the observations, and gives the response corresponding to the interval in which that difference lies, for example, ".7" or "very likely," or whatever the associated response is.

A calibration experiment consists of many such judgments. It generates distributions of the internal observation variable Y for correct and for incorrect alternatives, $f(Y|C)$ and $f(Y|\text{not } C)$, as in Figure 1a. The separation of these two distributions relative to their variances represents the discriminability of the observation process. The difference $y_C - y_{\text{not } C}$ between the value of Y for the correct alternative and that for the incorrect alternative is the value of the variable Z. The distribution $f(Z)$ is usually assumed to be normal. It will certainly be so if the distributions on Y are normal. It is shown in Figure 1b. The area of $f(Z)$ to the right of zero is the proportion of times the correct alternative produced the larger value of Y, that is, $p(C)$, the proportion of times the respondent made the correct choice. Since the actual scale is unknown, $f(Z)$ may be taken as a *unit* normal distribution with a mean that produces the experimentally observed value of $p(C)$ for the area above zero. The respondent, however, cannot know when Z is negative or when the incorrect alternative has the larger value of Y; the respondent sees only positive values of the difference Z. Thus, the decision variable X is the absolute value of Z.

The distribution $f(X)$, shown in Figure 1c, is the sum of two parts, the portion of $f(Z)$ below zero, with area $[1-p(C)]$, rotated about the vertical axis, and the portion of $f(Z)$ above zero, with area $p(C)$. The respondent is assumed to partition X into intervals by a set of cutoff values $\{x_i\}$, and to assign higher responses r_i to intervals farther out on X. When the observed value of $x = |y_C - y_{\text{not } C}|$ falls into the i^{th} region on X, the response r_i is given.[1]

The respondent experiences only $f(X)$ and the partition if there is no feedback. The experimenter, however, knowing which alternatives are correct, can calculate $p(C)$ and the proportions $p(r_i)$ and calibration value $p(C|r_i)$ of each response. Hence, the experimenter knows $f(Z)$ and can determine the corresponding partition (using various criteria for the fit). Figure 1d represents an experimenter's view of the model of the response process shown in Figure 1c. The proportion correct given

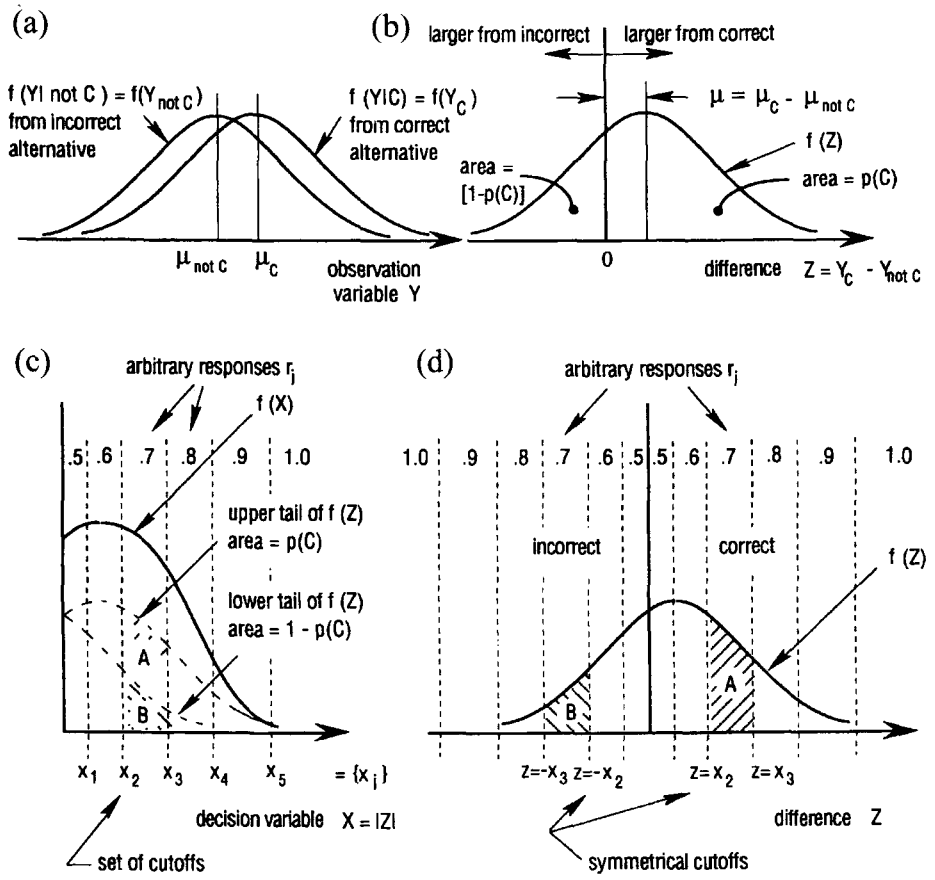

Figure 1. A graphical representation of the decision-variable partition model. (a) Distributions of the observations. (b) Distribution of the difference between the observations (generally assumed normal). (c) Distribution of the decision variable X, composed of components of $f(Z)$, as partitioned into response categories *by the respondent.* (d) Distribution of the difference between the observations Z, with partition, as inferred *by the experimenter,* $p(r = .7) = (A + B)$ and $p(C|r = .7) = A/(A + B)$.

a particular response $p(C|r_i)$ is just the ratio of the fraction of judgments that are correct when the decision variable X falls within the interval corresponding to response r_i to the total fraction of judgments that fall within that interval. In Figure 1d, $p(C|r_i = \text{".7"})$ is just $A/(A+B)$. A is the area in the interval corresponding to a response of ".7" under the component of $f(X)$ attributable to correct responses, and B is the area in that interval under the component attributable to incorrect responses. Furthermore, the sum $(A+B)$ is $p(r_i = \text{".7"})$, the proportion of responses of ".7."

Comparison With "Sensory Distance Theory"

The model presented by Björkman et al. (1993), which they call "sensory distance theory," is the same as this account of the two-alternative forced-choice, half-range, probability-correct task model originally presented in Ferrell and McGoey (1980); for the same data and criteria of fit, it produces precisely the same numerical results. The differences in presentation are due to the fact that the model discussed here is a special case of a more general one in Ferrell and McGoey (1980) that applies to a variety of task formats, not just to pair comparisons. Additionally, Björkman et al. (1993) approach the model from the standpoint of presignal detection-theory psychophysics, with appeal to Thurstone's discriminal difference rather than to an internal decision variable. Nevertheless, the models are the same and the equivalence is clearly evident from their description:

> Each comparison of the stimuli involves two discriminal processes [realizations of the internal observation variable] ... and, over trials, a resulting distribution of discriminal differences [the distribution of the (signed) decision variable].... Then, the proportion of correct responses is represented by the area to the right of zero, and the proportion of wrong responses is represented by the area to the left [as in Figure 1d].
>
> We now add the assumption that confidence is a monotonic increasing function of the difference between discriminal processes [this is a basic property of decision variables; the authors mean absolute value of the difference here].... Hence, the categories of confidence assessments x_t... are mapped into the continuum of sensory differences, with higher values of x_t going with larger differences [the partition]. (Björkman et al., 1993, p. 77)

Their Figure 1 illustrating the model, though less detailed, is the same as Figure 1d here.

The advantages of a signal detection theory approach are several. There is a great deal of directly relevant research and mathematical theory that one can draw upon, and there are well-defined and important linkages among sensitivity, decision criteria, and uncertainty. This paper will be concerned only with the model for two-alternative, half-range tasks, but other special cases of the decision-variable partition model have been found to fit the data for other question formats, as well (Ferrell, 1994a; Ferrell & McGoey, 1980; McClelland, Bolger, & Tonks, 1992).

Underconfidence Hypothesis

Underconfidence in calibration experiments with half-range tasks is defined as the responses r being less certain—that is, farther from 1 and closer to .5—than their corresponding proportion correct $p(C|r)$. Overconfidence is the reverse. Underconfident calibration curves thus tend to lie above the line of perfect calibration and overconfident ones to lie below it. Figure 1 shows that $p(C|r_1)$ for the lowest response category must be greater than .5 if that category width is >0 when $p(C) > .5$ because, in that case, the area corresponding to A is greater than that corresponding to B for a unimodal symmetric distribution. For this and the 1.0 response categories, the model predicts underconfidence.

Björkman et al. (1993) mistakenly generalize the underconfidence at the lowest response value to the entire calibration curve. They say, "We hypothesize that the imbalance between correct and wrong responses at $x_t = .5$ [$r_i = .5$ in present notation] extends to the higher categories of confidence with the consequence of an average underconfidence, $\bar{x} < \bar{c}$ [average $r < p(C)$ in present notation]" (p. 77). They also suppose that this underconfidence is independent of the proportion correct, saying, "The theory predicts underconfidence for all levels of $\bar{c}$ [$p(C)$] (with .5 and 1.0 as trivial exceptions)" (p. 79). This interpretation of the model is simply wrong. The "imbalance" in the lowest response category and the monotone increasing proportion correct of the model calibration curve do not together imply general underconfidence. Either under- or overconfidence can result, depending on the values of $p(C)$ and of the criterion cutoffs $\{x_i\}$. This is easily seen by exercising the model.[2]

Figure 2 gives the results of such an exercise. It shows, with solid lines, the calibration curves from the model for a fixed basic set of cutoffs and values of $p(C)$ ranging from .50 to .95. It also shows, with dashed lines, the curves from the model for a fixed $p(C) = .75$ and cutoffs that differ from the basic set by a proportional transformation, with proportions ranging from .5 (cutoffs 50% closer to zero on the decision axis) to 1.8 (80% farther from zero). These show how simple changes in cutoffs can compensate for changes in $P(C)$. The basic cutoffs {.35, .57, .74, .98, 1.2} have been arbitrarily chosen for illustration, but they fit the data for an actual experiment (Experiment 5 of Lichtenstein & Fischhoff, 1977).

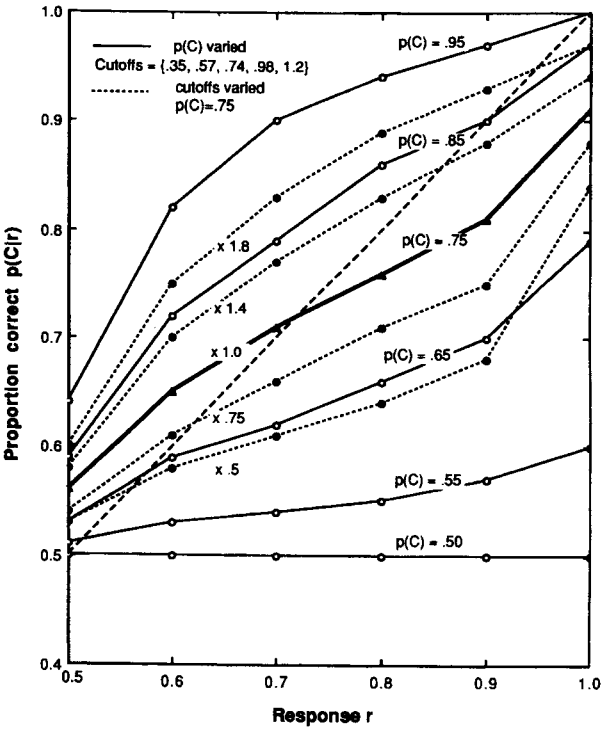

Figure 2. Model calibration changes due to: (a) changing $p(C)$ with fixed cutoffs (.35, .57, .74, .98, 1.2), $p(C)$ shown, and (b) changing cutoffs with fixed $p(C) = .75$, multiplicative constant shown.

It is clear from the figure that the two model parameters, the set of cutoffs $\{x_i\}$ and the proportion correct $p(C)$, enable the model to span the range from extreme over- to extreme underconfidence, the calibration curve shifting from far below to far above the diagonal that represents perfect calibration. The model predicts extreme overconfidence at very low discriminability, that is, $p(C|r) \rightarrow .5$ for all r, for *any* cutoffs when $p(C) \rightarrow .5$.

As an example of the calculations implied by the model, consider the point at $r_3 = .7$ on the curve for which $p(C) = .75$ and for which the cutoffs are the basic cutoffs of {.35, .57, .74, .98, 1.2}. Since $p(C) = .75$, the unit normal distribution $f(Z)$ in Figure 1d has a mean such that the area above zero is .75, that is, the mean μ is, thus, the value of the standard normal variate for which the cumulative distribution is .75. This is found from tables or by numerical approximation to be .675. The interval on Z corresponding to the response of ".7" when the choice is correct is the third interval above zero, since it is the third response in increasing order. The third interval is from .57 to .74. When the choice is not correct, the interval is the symmetrical one below zero, that is, from −.74 to −.57. In order to use the standard normal for determination of the areas corresponding to A and to B in Figure 1d, one must subtract the value of the mean μ from the values of Z to get the standard normal variate. Hence, the area corresponding to A is that between −.105 and .065 under the standard normal and the area corresponding to B is that between −1.415 and −1.245. Again

interpolating from tables or by numerical approximation, these areas are found to be $A = .0677$ and $B = .0280$, respectively. Thus, $p(C|r_3 = \text{".7"}) = A/(A+B) = .707$. The proportion of responses of ".7" is $p(r_3 = \text{".7"}) = \text{A+B} = .0957$. The latter value is not shown in Figure 2, but $p(r_i)$ is needed for a full description and is plotted in subsequent figures for comparison with data.

Changes in Calibration Due to the Proportion of Correct Choices $p(C)$

The model predicts that calibration can be changed by changing the proportion of correct choices $p(C)$. Figure 2 shows that if the cutoffs (and distribution type) remain fixed, changes in $p(C)$ can dramatically affect the calibration curve. If the task is made harder, the calibration curve shifts in the direction of overconfidence, that is, downward. If the task is made easier, the curve will shift in the direction of underconfidence, that is, upward.

Such a shift in the calibration curve with task difficulty for half-range tasks has been called the "hard–easy effect."[3] The effect is explained in the context of the model as the consequence of the respondent's failing to adjust response criteria appropriately, or doing so insufficiently, when there is a change in task difficulty, presumably because the respondent does not have the necessary information or lacks motivation. This explanation was first presented in Ferrell and McGoey (1980).

Björkman et al. (1993) claim that the shift is not observed with sensory discrimination judgments, and assert that the model supports this claim. They say, "Psychophysical discrimination does not exhibit any hard–easy effect, and it shouldn't. The theory predicts underconfidence for all levels of $\bar{c}$ [$p(C)$] (with .5 and 1.0 as trivial exceptions)" (p. 79). They are wrong on both counts. The model allows such a shift, as can be seen in Figure 2. Moreover, in an experiment by Keren (1988), which they cite as supporting a general underconfidence bias for sensory judgment, there is a clear hard–easy effect, in good agreement with the model, as demonstrated below.

The task used by Keren (1988) involved visual discrimination between left- and right-facing Landolt rings, annuli with a small gap in the lower left or right quadrant. Subjects attempted to identify the direction of the gap in rings presented briefly, one at a time, and indicated their confidence that they had identified correctly. Although there are two alternative responses and confidence is on the interval 0.5 to 1.0, it is not a paired-comparison judgment, since there is only one observation on each trial. In this case, the decision variable that is partitioned in the model is the sensory observation variable, itself. When the alternatives are treated equally in the experimental design (i.e., presented equally often and with the same reward/penalty structure, as here), the partition is assumed to be symmetrical about a point halfway between the normal distributions for the two types of observation, and the calculations and the model properties are indistinguishable from those of the version of the model that applies to pair comparisons (Smith & Ferrell, 1983).

Two gap sizes were used equally often and randomly intermixed. The small gap produced $p(C) = .67$ and the large gap, $p(C) = .79$. The calibration results are shown in Figure 3 along with the model fitted independently[4] to the results for each gap size. The harder task shows a clear shift toward overconfidence in both the experimental results and the fitted model. Since the gap sizes were randomly intermixed in the experiment and the stimuli were otherwise the same, it is plausible that the cutoffs would be the same for each type of stimulus, that is, there would be no cue that would enable respondents to change cutoffs when a stimulus with a different gap size was presented. The cutoffs obtained from independently fitting the model in the two cases are {.3, .78, 1.36, 1.92, 2.23} for the large gap and {.36, .89, 1.44, 1.88, 2.20} for the small. These are not different by a paired t test ($p = .298$). Hence, the shift from under- toward overconfidence when the gap is reduced can be attributed to the cutoffs being held constant when the task is changed in difficulty on randomly chosen trials. Thus, contrary to Björkman et al. (1993), a hard–easy effect has been observed with a purely sensory task, and it is explained by the model rather than prohibited by it, according to

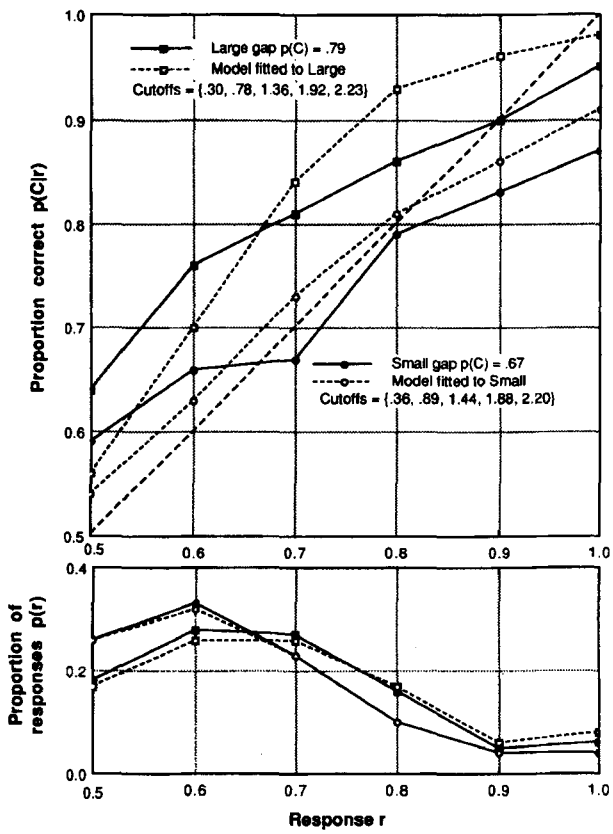

Figure 3. Calibration and response proportion curves for Keren's (1988) Landolt-ring experiment.

their interpretation. Subsequently, a hard–easy effect has been clearly demonstrated in experiments on comparison of linear extent by Baranski and Petrusic (1994).

Björkman et al.'s (1993) calibration results from their Experiment 1 show a shift that might be interpreted as due to the hard–easy effect for comparison of weights but not for comparison of rectangles. However, it is impossible to interpret their results in terms of the model, because they have defined their *hard* and *easy* conditions and analyzed their data in a way that would tend to obscure any hard–easy effect, and they do not report the details that would permit a different analysis.

Consider their rectangle task. It consisted of presentations of a stimulus with a standard. There were four pairs of stimuli, differing from the standard by progressively greater amounts. One stimulus of a pair was less and one was greater than the standard by the same amount. The *hard* condition and the *easy* condition both had three of the pairs in common, and differed with respect to only one stimulus pair each, a pair very close to the standard for the *hard* condition and one very different for the *easy*. For each condition, the authors fitted the model separately to data for each of the stimulus pairs, and then cumulated the results over response categories across stimulus pairs. They report only the final results. Thus, each model calibration curve incorporates data from four realizations of the model, having, presumably, different parameters.

It would have been more in keeping with the model to treat, within the *hard* or *easy* condition, each of the stimulus pairs (differing in objective difficulty) as constituting a subtask and to analyze any effect of difficulty at that subtask level rather than pooling the results over the rather arbitrarily defined conditions. If this were done, one might predict a within-task hard–easy effect with calibration of confidence judgments for the rectangles that are harder to discriminate, being less underconfident than for those easier to discriminate.

Moreover, the *hard* and *easy* conditions might well have been perceived differently, since the hard part of the *hard* condition is quite hard and the easy part of the *easy* condition is quite easy. As a result, one might expect subjects to adopt somewhat different criteria for the two conditions. The analysis fails to make use of the model to examine such questions.

It should be noted that the model calibration curves in Figure 3, fitted by the maximum likelihood method, systematically deviate from the data. This suggests that the distributions on the decision variable are not normal, as is assumed by the fitting process. It points up the fact that the normal assumption is really only an approximation adopted because it so often works. But the assumption of normality of the conditional distributions on the decision variable is not invariably appropriate (Egan, 1975) and is not a fundamental feature of the model. Hellström (1993) shows that sensory data that are usually assumed normal for each individual can be significantly nonnormal when pooled over subjects.

It should be noted also that when the model is applied to data sets of typical size, as above, using the maximum likelihood method, the results are frequently not statistically indistinguishable from the data at the confidence level of .05. The model with the normal distribution and $p(C)$ specified is quite demanding. Twelve data points must be fitted having an elaborate pattern of interdependence among them dictated by the distribution form and by the order of the cutoffs. Björkman et al.'s (1993) statistically significant results may appear to be an exception. However, by fitting the response proportions $p(r_i)$ exactly, they have assumed that the cutoffs estimated from the data are specified as part of the "theory." This substantially improves the possibility of finding no significant statistical difference between the model and the data. From a psychological standpoint, however, the cutoffs are an important dependent variable. Nevertheless, the quality of fit to any particular data set is of much less importance than the overall pattern of coherent representation and prediction that the model shows with a variety of data sets, and the fact that the more general model extends to and fits full-range judgments.

Confidence in Cognitive as Compared With Sensory Judgment

Björkman et al. (1993) conclude that experimental evidence "suggests a different nature of confidence in sensory judgments, relative to cognitive judgments" (p. 81). In support of this, they cite (1) the frequent finding of overconfidence with general-knowledge questions, comparing it to their findings of underconfidence for sensory comparisons, and (2) the success of the model for sensory judgments discussed here and the success with cognitive judgments of what has been called (McClelland & Bolger, 1994) the "ecological model" (Gigerenzer, Hoffrage, & Kleinbolting, 1991; Juslin, 1994). It is notable that they did not cite any particular inability of the model discussed here to deal with cognitive judgments.

This journal is not the appropriate forum for discussing the merits of models of cognitive judgment, but it must be said that the general model of calibration presented in Ferrell and McGoey (1980), and the portion of it examined here, accurately describes calibration of subjective probabilities both for sensory and for a variety of cognitive judgments. It fits half- and full-range judgments and explains or accommodates essentially all the robust experimental findings for calibration, including the hard–easy effect described above and the base-rate effect (Ferrell, 1994b; Lichtenstein et al., 1982; Smith & Ferrell, 1983). (It does not, however, purport to describe subjective probability judgments that are the result of mental calculations or operations on numbers, but only those that can be considered to be based on the partitioning of an otherwise unscaled internal variable.)

An example of the model's application to cognitive judgment and its ability to explain observed differences in calibration in the same manner as with sensory judgment is shown in Figure 4. The data are from Juslin

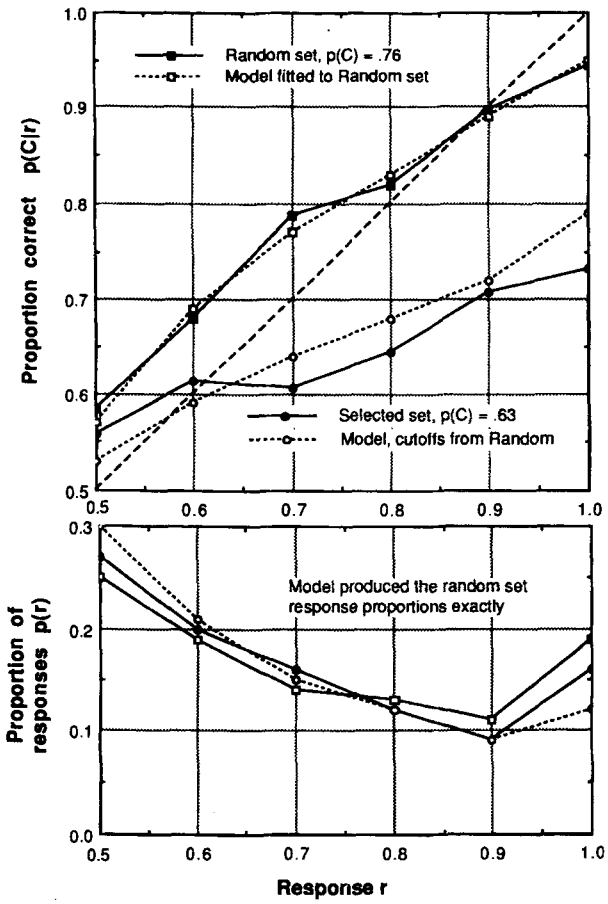

Figure 4. Example of the model used with cognitive judgments. Data from Juslin (1994). Cutoffs = .41, .73, 1.01, 1.31, 1.62. (From "The Overconfidence Phenomenon as a Consequence of Informal Experimenter-Guided Selection of Almanac Items" by P. Juslin, 1994, *Organizational Behavior & Human Decision Processes, 57*, pp. 226-245. Copyright 1994 by publisher. Reprinted by permission.)

(1994)[5] and represent the calibration results from two-alternative forced-choice questions about pairs of countries. For the *random* set, sampling of questions was such that every pair of countries was equally likely to be included. For the *selected* set, questions were selected (by subjects other than those answering the questions) as if for a test of knowledge. Juslin (1994) concluded that the poor calibration and overconfidence of the *selected* set were artifacts of the selection process.

The model was fitted to the random set, with results as shown in the figure. On the assumption that the experiments and subjects were sufficiently similar, the resulting cutoffs were then used with the value of $p(C) = .63$ from the *selected* set to predict the calibration for it. As can be seen, the fit to both the calibration values and the response proportions is quite good, considering that only one parameter was estimated from the data for the 12 graph points. Consequently, the difference in calibration for the two sets of questions is well explained by the model as the result of the greater difficulty of the *selected* set. It is an example of the same "hard–easy" effect exhibited in Keren's (1988) Landolt-ring experiment in Figure 3. Without calibration feedback or cues to difficulty, subjects would not be expected to be able to adjust their response criteria to match question difficulty. Although the model fits the data for the *random* set well, it does not predict their good calibration. Juslin attributes the good calibration to the unbiased selection of questions and to subjects knowing the numerical validities of the cues they use to decide the truth of uncertain propositions. However, good calibration is commonly found for two-alternative forced-choice cognitive questions when the proportion correct $p(C)$ is about .75 even when they are not randomly selected. Why this is so is not clear.

The "Underconfidence Phenomenon" Reconsidered in Light of the Model

Underconfidence with judgments of correctness of sensory discrimination is a robust phenomenon, according to Björkman et al. (1993). It may, indeed, be common, but it is certainly not universal. Overconfidence has been reported by a number of authors for half-range perceptual tasks that have no evident cognitive content, for example, identification of one of five words in noise (Clarke, 1960, as reported in Lichtenstein et al., 1982), discrimination of the area of figures (Lichtenstein & Fischhoff, 1980), discrimination of American and European handwriting (Lichtenstein & Fischhoff, 1977), and comparison of visual extent (Baranski & Petrusic, 1994). In all of these cases, the hard–easy effect appears to have been present, so that the overconfidence would be a result of the interaction of response criteria $\{x_i\}$ and difficulty $p(C)$.

In the context of the model, response criteria are crucial to whether there will be actual under- or overconfidence for a given task. The ability to change response criteria systematically is well documented in the signal-detection literature (Decker & Pollack, 1958; Green & Swets, 1974). Indeed, the creation of receiver operating characteristic (ROC) curves using a series of yes–no experiments, each of which determines a different point on the curve, has routinely been done by changing the subjects' reward/penalty function or by changing the proportion of signal trials. Even instructions to adopt a more or a less strict criterion have been found to be successful.

The ability to appropriately change a whole *set* of criteria, not just one, might be questioned, but there is good evidence for it, also, in both the signal-detection literature (Decker & Pollack, 1958) and the calibration literature. Lichtenstein and Fischhoff (1980), using the two-alternative forced-choice paradigm with the sensory task of judging the area of irregular shapes, found "that a single session of 200 items followed by intensive performance feedback [calibration, not outcome] is sufficient to teach people who are not initially well calibrated to be well calibrated. This improvement occurs without the subjects ever learning the true answers to any items" (pp. 167-168). Since overconfidence was reduced at the

same task without any improvement in $p(C)$, the criteria, in terms of the model, must have changed.

Individuals may respond to cues that suggest the degree of difficulty of the task by making their response criteria, the cutoffs, generally somewhat more strict (moving them outward) for tasks that appear to be harder and more lax (moving them inward) for those that seem easier. In terms of the model calibration results shown in Figure 2, this would be an attempt to have the shift of the calibration curve caused by cutoff change compensate for the shift caused by a change in task difficulty $p(C)$. This possibility seems highly likely for two reasons. First, the cutoffs are, in principle at least, quite arbitrary and independent of the task. Second, Figure 2 shows that a very simple one-parameter movement of cutoffs, that is, a percentage change, can shift the calibration in a manner very similar to that caused by a change in $p(C)$. It seems unlikely, however, that there would be a very precise or universal relationship between perceived task difficulty and the location of cutoffs. It is more likely that a movement of cutoffs would be relative to some initial, perhaps quite inappropriate, default location, possibly even making calibration worse.

It is plausible, in addition, that cognitive judgment tasks are reacted to as if they are less difficult than sensory ones, whether this is consciously experienced or not, in view of the evidence for greater certainty in matters over which one seems to have control (Heath & Tversky, 1991). As a result, one might expect to find default response criteria to be systematically more strict for sensory tasks than for cognitive tasks, leading more often to underconfidence.

Conclusions

It has been pointed out that Björkman et al. (1993) have not presented a new model of calibration, but a portion of the more general decision-variable partition model that was originally proposed in Ferrell and McGoey (1980). More important, analysis shows that they have misinterpreted that model. Contrary to Björkman et al. (1993), (1) the model does *not* predict general underconfidence, (2) the hard–easy effect is quite possible, according to the model, and has, in fact, been observed with sensory discriminations, and (3) the model fits not only sensory judgments, but cognitive ones as well. Moreover, the frequent, but not universal, observation of overconfidence with cognitive judgments and underconfidence with sensory judgments does not need two models to explain it; it can be accommodated as the result of a common causal mechanism within the context of the decision-variable partition model.

REFERENCES

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

BJÖRKMAN, M., JUSLIN, P., & WINMAN, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics*, **54**, 75-81.

CLARKE, F. R. (1960). Confidence ratings, second choice responses, and confusion matrices in intelligibility tests. *Journal of the Acoustical Society of America*, **32**, 35-46.

DECKER, L., & POLLACK, I. (1958). Confidence ratings and message reception for filtered speech. *Journal of the Acoustical Society of America*, **30**, 432-434.

EGAN, J. P. (1975). *Signal detection theory and ROC-analysis*. New York: Academic Press.

EGAN, J. P., SCHULMAN, A. F., & GREENBERG, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America*, **31**, 768-773.

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. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 411-451). New York: Wiley.

FERRELL, W. R., & MCGOEY, P. J. (1978). *A model of calibration for subjective probabilities* (Human Factors & Man-Machine Systems Laboratory Report). Tucson: University of Arizona, Systems & Industrial Engineering Dept.

FERRELL, W. R., & MCGOEY, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior & Human Performance*, **26**, 32-53.

FERRELL, W. R., & REHM, K. (1980). A model of subjective probabilities from small groups. In *Proceedings of the Sixteenth Annual Conference on Manual Control* (pp. 271-284). Cambridge, MA: MIT Press.

GIGERENZER, G., HOFFRAGE, U., & KLEINBOLTING, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, **98**, 506-528.

GREEN, D. M., & SWETS, J. A. (1974). *Signal detection theory and psychophysics* (2nd ed.). New York: Wiley.

HEATH, C., & TVERSKY, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk & Uncertainty*, **4**, 5-28.

HELLSTRÖM, Å. (1993). The normal distribution in scaling subjective stimulus differences: Less "normal" than we think? *Perception & Psychophysics*, **54**, 82-92.

JUSLIN, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior & Human Decision Processes*, **57**, 226-245.

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.

LICHTENSTEIN, S., & FISCHHOFF, B. (1977). Do those who know also know more about how much they know? *Organizational Behavior & Human Performance*, **20**, 159-183.

LICHTENSTEIN, S., & FISCHHOFF, B. (1980). Training for calibration. *Organizational Behavior & Human Performance*, **26**, 149-171.

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

MCCLELLAND, A. G. R., & BOLGER, F. (1994). The calibration of subjective probabilities: Theories and models 1980-1994. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453-482). New York: Wiley.

MCCLELLAND, A. G. R., BOLGER, F., & TONKS, E. (1992, January). The effects of discriminability and base rate on calibration of probabilities in a perceptual discrimination task. In N. Harvey & P. Ayton (Chairs), *Judgment and decision making*. Symposium conducted at the meeting of the Experimental Psychology Society, London.

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* (pp. 469-488). Amsterdam: North-Holland.

SWETS, J. A., TANNER, W. P., JR., & BIRDSALL, T. (1961). Decision processes in perception. *Psychological Review*, **68**, 301-340.

NOTES

1. These components of $f(X)$, representing correct and incorrect responses, are more formally described by the conditional distributions on X, $f(X|C)$ and $f(X|\text{not}C)$, as in Ferrell and McGoey (1980). This leads to additional mathematical terms, but has the advantage of making it easier to express the general model which does not depend upon the decision variable's being a difference of two values.

2. A HyperCard stack for Macintosh computers that implements the model for two-alternative forced-choice tasks is available from the author (please send a diskette). It allows input of the cutoffs, graphically shows them in the manner of Figure 1c, and plots the resulting calibration curve and response proportions.

3. This effect is also observed with full-range tasks and is predicted by the decision-variable partition model in that case as well (Ferrell, 1994; Ferrell & McGoey, 1980).

4. The fitting was done by the maximum-likelihood technique using the program ROCFIT with input modified to accommodate the 2AFC model. The program was obtained by courtesy of Charles Metz, Department of Radiology, University of Chicago, in whose laboratory it was developed.

5. P. Juslin has very kindly supplied the response proportions, enabling the model to be fitted.

(Manuscript received July 19, 1993;
revision accepted for publication July 5, 1994.)