Psychological Review
1994, Vol. 101, No. 3, 519-527

0033-295X/94/$3.00

Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processes

Ido Erev, Thomas S. Wallsten, and David V. Budescu

Two empirical judgment phenomena appear to contradict each other. In the revision-of-opinion literature, subjective probability (SP) judgments have been analyzed as a function of objective probability (OP) and generally have been found to be conservative, that is, to represent underconfidence. In the calibration literature, analyses of OP (operationalized as relative frequency correct) as a function of SP have led to the opposite conclusion, that judgment is generally overconfident. The authors reanalyze 3 studies and show that both results can be obtained from the same set of data, depending on the method of analysis. The simultaneous effects are then generated and factors influencing them are explored by means of a model that instantiates a very general theory of how SP estimates arise from true judgments perturbed by random error. Theoretical and practical implications of the work are discussed.

The focus of this article is an apparent paradox in the domain of human judgment. Two well-documented and robust empirical phenomena appear to be in direct contradiction: On the one hand, a large body of literature produced primarily in the 1960s demonstrates that when subjects are required to estimate probabilities of hypotheses following the observation of data, their estimates generally are less extreme than values properly calculated by means of Bayes rule. (Reviews of this *revision-of-opinion* literature have been provided by Edwards, 1968; Rapoport & Wallsten, 1972; Slovic & Lichtenstein, 1971; and more recently by Fischhoff & Beyth-Marom, 1983.) The locus of this underconfidence (or conservatism, as the phenomenon was called) was never determined, and for a variety of reasons "this line of research was quietly abandoned" (Fischhoff & Beyth-Marom, 1983, p. 248). Among the criticisms leveled at the paradigm was that the tasks are too unfamiliar and complex (Pitz, Downing, & Reinhold, 1967) or simply too far removed from the real world (Winkler & Murphy, 1973) to yield generalizable results.

Partially in response to such criticisms, and without questioning the existence of underconfidence or conservatism, a new paradigm was developed in which subjects are asked to estimate the probabilities that statements, answers to questions, or forecasts are correct. Generally, the assessments are more extreme than the associated relative frequencies correct, suggesting that subjects are overconfident in their judgments. (For reviews of this literature on *calibration*, see Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1982; Wallsten & Budescu, 1983; Yates, 1990.) This newer paradigm is not without its critics. For example, Gigerenzer (1991) and Gigerenzer, Hoffrage, and Kleinbölting (1991) have criticized the calibration research by arguing, among other things, that probabilities do not apply to unique events.

Nevertheless, the results from the two paradigms seem to be in direct contradiction. An easy resolution of the conflict would be to claim that underconfidence[1] occurs in revision-of-opinion tasks and overconfidence occurs in forecasting or general knowledge cases, but that distinction will not do. For one thing, the opposite results also have been found under some circumstances in each paradigm. For another, the two tasks are often difficult to distinguish. To illustrate that fact, consider the following hypothetical scenario. A businessman (DM, for decision maker) must decide whether to invest a certain amount of money to achieve a desirable deal. An expert assesses the chances of getting the contract without the investment as .25 and with the investment as .80. Under the assumption that her assessments are accurate, the DM is indifferent between investing and not investing. While discussing the problem with two consultants, the DM is reminded that subjective probabilities (SPs) are not always equivalent to objective probabilities (OPs).

In particular, one consultant says, "The expert gave her judgment after considering the available data. Bayes rule provides the normative way to revise opinion in such cases, and considerable research has shown subjectively estimated probabilities to be insufficiently extreme relative to this standard. Generally,

Ido Erev, Department of Psychology, The Technion, Israel Institute of Technology, Haifa, Israel; Thomas S. Wallsten, Department of Psychology, University of North Carolina at Chapel Hill; David V. Budescu, Department of Psychology, University of Illinois.

The order of authorship is arbitrary. The three authors made equal contribution to the work, which was supported by Grants BNS-8908554 and SBR-9222159 from the U.S. National Science Foundation and by the Technion Vice President for Research Fund–New York Metropolitan Fund.

We wish to thank Barbara Mellers for conversations that stimulated some of the developments in this article and Ayala Cohen, Robyn Dawes, Claudia González-Vallejo, Daniel Kahneman, David Navon, Tom Nelson, Amos Tversky, and Joseph Young for helpful comments on previous drafts or following presentations of this research.

Correspondence concerning this article should be addressed to Thomas S. Wallsten, Department of Psychology, University of North Carolina, Chapel Hill, North Carolina 27599-3270. Electronic mail may be sent to tom_wallsten@unc.edu.

[1] To maintain parallel language, we consistently use the term *underconfidence* instead of *conservatism*.

when subjects assess probabilities to be .80, the OPs are around .95. When subjects judge probabilities to be .25, the objective estimates are in the neighborhood of .10. Thus, given that the best probability estimates are probably more extreme than the expert's, it appears that investing may be the better decision for you."

The second consultant, however, says that, "In studies focusing on situations similar to the present problem, subjects have been asked questions about real-world events and have been required to judge the probabilities that their answers were correct. It was found that people are poorly calibrated and generally too extreme in their estimates. That is, when subjects said that the probability was .80, the actual relative frequency of correct judgments was only about .65. When they assessed the probability to be .25, the relevant relative frequency was as high as .40. Thus, given that the best probability estimates are probably not as extreme as the expert's, it seems that not investing is your best course."

The two recommendations, of course, are contradictory, and the literature provides no clear guidance as to which (if either) might be correct. Is this a case of forecasting, in which judges are generally overconfident, or of opinion revision, in which they are generally underconfident? Of course it is both, as occurs in most real situations where people consider data for the purpose of answering a question or providing a forecast. Thus, a theoretical framework is needed to understand the conditions leading to under- and overconfidence in judgment, and none is currently available. We propose an approach that provides guidance to the businessman and more generally provides a broad framework within which to theorize about probability judgments and to consider issues of under- and overconfidence in judgment. To anticipate our results, we propose a class of models that explicitly incorporate error in the judgment or response process and that can handle the full range of results reviewed above, depending on the method of data analysis, which it turns out is crucial.

We begin with the observation that the revision-of-opinion literature, where the underconfidence is the dominant finding, has focused on paradigms in which prior probabilities of hypotheses and conditional probabilities of data are well-defined stochastic properties of the environment. Uncertainty, therefore, is external to the subject (Budescu & Wallsten, 1987; Howell, 1972; Howell & Burnett, 1978; Kahneman & Tversky, 1982) and naturally quantified as posterior probabilities. In contrast, the calibration literature, where overconfidence is the dominant finding, has focused on paradigms in which uncertainty is due primarily to lack of knowledge rather than to well-defined environmental factors. In these cases, uncertainty is internal to the subject and not independently measurable.

This difference in paradigms has two potential implications. The observed pattern of data may be due to a fundamental difference in how humans handle internal and external uncertainty, perhaps along the lines suggested by Griffin and Tversky (1992). However, it is also possible that the pattern stems from the types of data analysis that naturally accompany each paradigm. Because OPs are independently defined in the revision-of-opinion case, it has been customary to investigate SP as a function of OP. That is, inferences have been drawn about the *relationship*, $SP = f(OP)$. In contrast, because OP is not independently definable in the calibration paradigm, investigators have looked at the inverse relationship, $OP = g(SP)$.[2] The conflict arises because $f \neq g^{-1}$; rather, both f and g are regressive.

It is well known that when two variables, X and Y, are not perfectly linearly related to each other, the prediction of each from the other is generally regressive to the mean. However, this fact does not mean that the conflict is purely a statistical artifact. Although it might be argued that SP is the same variable in both paradigms, namely subjects' confidence in an alternative, OP is surely not the same in both cases. In the revision-of-opinion paradigm, OP is defined independently of any judgment, whereas in the calibration paradigm, it is a relative frequency measure of equivalence classes created by the subject.

Thus, it is an empirical question whether the conflicting patterns of results are due to the methods of analysis and corresponding definitions of OP or to fundamental differences in how subjects handle internal and external sources of uncertainty. To address the issue, it is necessary to apply both types of analyses to sets of judgments involving internal uncertainty, external uncertainty, or both and where the two definitions of OP can in principle coincide. If the conflict is due largely to information-processing factors, then the method of analysis should not have a large impact on the inference drawn from a particular data set. Contrariwise, if the conclusion about over- versus underconfidence depends on the analysis across data sets, then we must develop a theoretical framework that takes this complication into account.

In the next section, we reanalyze three data sets and show simultaneous over- and underconfidence of similar degrees, depending on the analysis. These data sets vary in whether the uncertainty is internal or external, but all have natural ways to define OP. We then present a very general model, incorporating the assumption that observed responses are a function of true judgment and error, and illustrate it with a specific instantiation, illustrative of others that are possible, that can yield the full pattern of results. Finally, in the Discussion section, we consider theoretical and practical implications of the work. We advise our poor businessman and in the process suggest a theoretical perspective on judgment research that distinguishes and accommodates basic theoretical and real-world operational concerns.

Probability Judgments That Appear Both Over- and Underconfident

Assessments in Probabilistic Video Game

The main task of subjects in the study by Erev and Wallsten (1993) was to choose among gambles based on events in a video game. In one of the conditions, the subjects (60 University of North Carolina [UNC] students) were also asked to assess the probabilities of winning the gambles. The video display consisted of vertical tracks, each with an object at its left end. Within each track were barriers that continuously opened and

[2] We must define OP carefully here and do so in a relative frequency sense. Specifically, we assume that subjects categorize events according to their confidence that the events are true or will occur and attach a different response (SP) to each category. OP is the relative frequency of true or occurring events in each category.

closed. During the game, each object on the left was released and either reached the right edge of the track, in which case the gamble was won, or crashed into a barrier and exploded. The probability that each object would reach the right edge was predetermined and visually displayed as the proportion of time that the barrier's opening was wider than the object.

The present analysis uses the probability assessments of 36 video game events by each of the subjects. They were instructed on how to gauge the probabilities that objects would pass the barriers. That the subjects seemed to understand is attested to by the relative accuracy of their assessments, which had a median correlation over subjects of .75 with the OPs. Figure 1 presents mean SP estimates over the 60 subjects as a function of the OPs for the 36 events.

The following approach was taken to provide parallel analyses of the two types. First, corresponding to analyses when OP is not defined, each SP judgment was categorized according to its nearest multiple of .10. (In other words, 11 response categories were formed, [0, .05), [.05, .15), . . . , [.85, .95), [.95, 1.0].) Mean OP (corresponding to expected relative frequency correct) and mean SP were calculated for the event–response pairs in each response category. These results are shown as the S curve in Figure 2, so named because data are grouped according to the SPs. Then, corresponding to analyses when OP is defined, each OP was categorized to its nearest multiple of .10, following which mean OP and mean SP were calculated for the event–response pairs in each of the 11 OP categories. These results are shown as the O curve in Figure 2, again, so named because data were grouped according to the OPs.

Figure 2 is oriented in the manner common for studies in which OP is defined, and the O curve is easily recognized as the signature for conservative probability estimation, or as we are calling it here, underconfidence. The orientation is opposite to that common in calibration research, but making the adjustment, the S curve demonstrates overconfidence. Thus, these data imply underconfidence when analyzed as is standard in one case and overconfidence when analyzed as is standard in the other.[3]

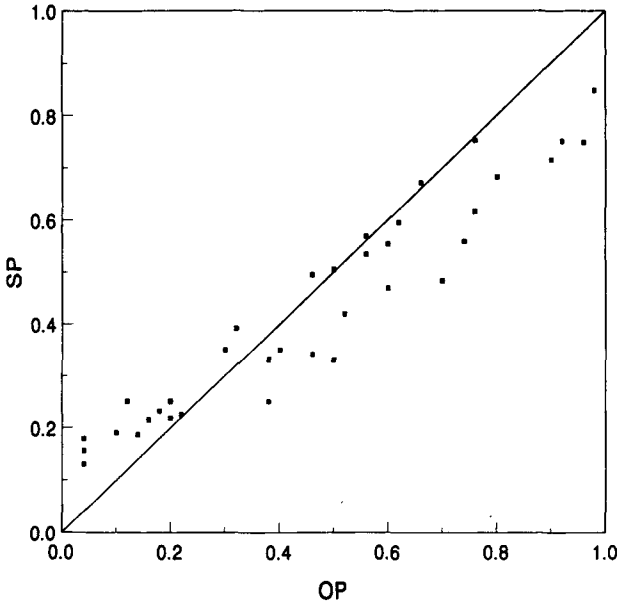

Figure 1. Mean subjective probability (SP) as a function of objective probability (OP) for the 36 events (Erev & Wallsten, 1993).

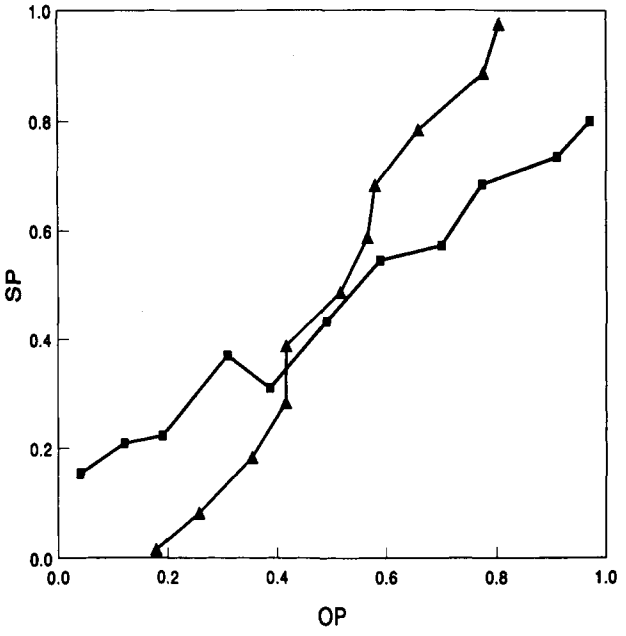

Figure 2. Mean objective probability (OP) as a function of subjective probability (SP; S curve denoted by ▲) and mean SP as a function of OP (O curve denoted by ■) for the Erev and Wallsten (1993) study.

It is important to examine whether simultaneous under- and overconfidence also can be detected at the individual level. To achieve this goal, two scores were calculated for each subject to measure the average deviation associated with each analysis. $CONF_O$, an index of relative confidence when conditioning on OP, is the average signed deviation between each mean SP and the corresponding mean true OP. That is, for n events and $i = 1, \ldots, n$,

$$CONF_O = \frac{1}{n}\left[\sum_{OP_i < .5} (OP_i - SP_i) + \sum_{OP_i > .5} (SP_i - OP_i)\right].$$

Events with OP = .5 are excluded from this score. Positive values of $CONF_O$ indicate overconfidence and negative values indicate the opposite, or underconfidence. In a similar manner, $CONF_S$ is an index of relative confidence when conditioning on SP. That is,

$$CONF_S = \frac{1}{n}\left[\sum_{SP_i < .5} (OP_i - SP_i) + \sum_{SP_i > .5} (SP_i - OP_i)\right].$$

[3] Actually, when OP is defined, as in Bayesian revision-of-opinion studies, it is common to average SP for each OP value, not for an interval of values. In fact, that is what was plotted in Figure 1, where the general pattern of underconfidence can be discerned. We focus on OP intervals to maintain parallelism between the two types of analysis.

As before, positive values indicate overconfidence and negative values indicate the opposite, or underconfidence. Note that $CONF_O$ and $CONF_S$ differ not only in how the summations are defined but also in the values that are summed due to different conditionings in the two cases.

In the present data set, 54 of the 60 subjects (90%) had negative $CONF_O$ scores, appearing underconfident when the assessments are grouped by OPs. In addition, 54 of the subjects (90%) had positive $CONF_S$ scores, appearing overconfident when the assessments were grouped by the subjective responses. These results are consistent with the aggregated data analysis and demonstrate that simultaneous under- and overconfidence can be observed on a within-subject basis.

Obviously, in a substantive sense, the same assessments cannot be both under- and overconfident. Possibly, we have done nothing more than demonstrate statistical regression in a single errorful data set from a study that is somewhat different from both the usual calibration and the usual revision-of-opinion experiments. However, as the next two data sets show, the identical results obtain in these standard paradigms.

Assessments of Future Basketball Events

Ronis and Yates (1987) found subjects to be overconfident when estimating probabilities of future basketball events. In their study, as in many others supporting the overconfidence hypothesis, it was impossible to group the data based on independent estimates of OP because independent estimates were not available. Consider a variant of Ronis and Yates's study in which subjects are asked to estimate the probability of a repeated event in a basketball game (e.g., the probability that event *E* will occur in one game to be randomly chosen out of the next *n* games). Here OP can be estimated as the proportion of games in which the event occurred.

In an experiment by Erev, Bornstein, and Wallsten (1993), 60 subjects (UNC students) had to choose among gambles. Each gamble promised a certain amount of money if a particular event were to occur in a game randomly selected from the list of the nine forthcoming basketball games involving UNC. For example, one of the events was "Pete Chilcutt will score more than 15 points." Lists of the nine games were provided, and the subjects were asked to give the relevant SPs. After the nine games were played, the OPs were estimated by the proportion of games in which each event occurred. The median correlation between the subjects' estimates and the actual proportions was .53.

In Figure 3, we present the mean SP as a function of mean OP for event–response pairs grouped into categories according to nearest multiples of .1 by the subjective estimates (the S curve) and by the actual proportions of occurrence (the O curve). The S curve shows the familiar pattern of overconfidence. For instance, mean assessments close to 0 correspond to an average occurrence rate of .13. As in the video game data, the O curve can be interpreted as indicative of the opposite trend. Events that occurred in about .8 (.77, to be exact) of the games were assessed with a mean SP of .61.

The present analysis seriously questions the existence of an overconfidence tendency per se in judging the probabilities of future basketball events. As in the video game environment, the grouping of the data determines the generalization that is inferred from them. We look now into the possibility of finding evidence of overconfidence in a task that typically gave rise to underconfidence.

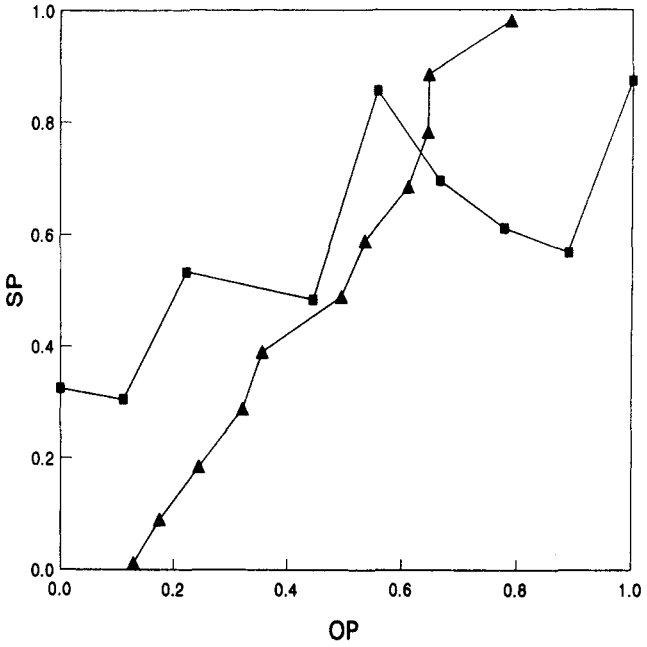

Figure 3. Mean objective probability (OP) as a function of subjective probability (SP; S curve denoted by ▲) and mean SP as a function of OP (O curve denoted by ■) for the Erev, Bornstein, and Wallsten (1993) study.

Assessments in a Two-Hypothesis Revision-of-Opinion Task

Many studies have compared SP judgments with the predictions of Bayes rule in a paradigm involving the sampling of chips from one of two urns with specified compositions and well-defined prior probabilities. Assessments could easily have been grouped on the basis of either SPs or OPs. Typically (e.g., Phillips & Edwards, 1966; Rapoport, Wallsten, Erev, & Cohen, 1990) the subjects' responses (SP) were grouped on the basis of the levels of the independent variable that was manipulated by the experimenter (OP). Whereas this convention is natural, the analyses presented above suggest that it can influence the conclusion drawn from the results.

We reanalyzed that portion of the Rapoport et al. (1990) data in which subjects gave numerical (rather than verbal) assessments, grouping the data each way. Rather than selecting an urn following the observation of a sample of chips, the 24 subjects assessed the probability that the sample came from urn A rather than urn B. Thus, a full [0–1.0] rather than a half [.5–1.0] response scale was used. In Figure 4, we compare the S- and the O-curve presentations of the data. As in the previous data sets, the S curve appears to reflect overconfidence. For example, when the subjects assessed the probability to be close to 0 (mean SP of .01), the mean OP was .15. Simultaneously, the O curve is

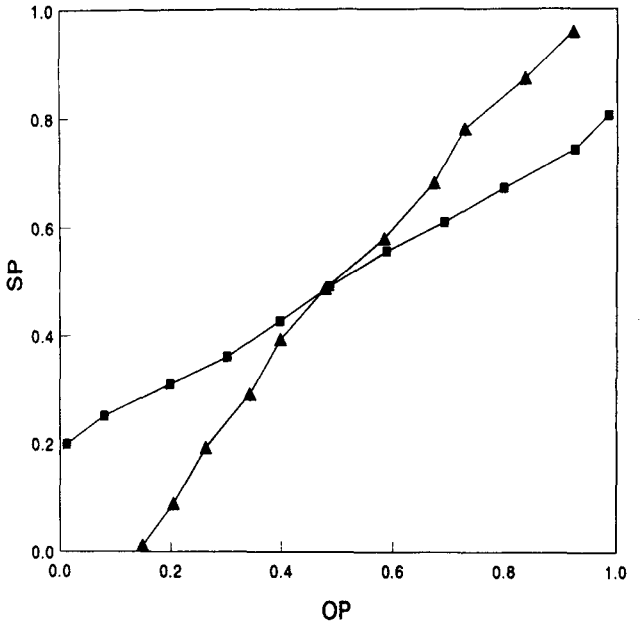

Figure 4. Mean objective probability (OP) as a function of subjective probability (SP; S curve denoted by ▲) and mean SP as a function of OP (O curve denoted by ■) for the Rapoport, Wallsten, Erev, and Cohen (1990) study.

consistent with the conservatism hypothesis of underconfidence. For instance, when mean OP is .01, mean SP is .20.

Thus, the analysis of these data from a standard revision-of-opinion experiment agrees with the previous analyses. The conclusion seems inevitable that the way the data are conditioned determines the conclusion that is drawn from them.

Assessments in Blackjack

Apparently without recognizing the implications, Wagenaar and Keren (1985) also presented S curves and O curves that showed the same patterns observed in our analyses. They were interested in comparing probability judgments of blackjack experts, statisticians, and a control group with regards to the chances of various events occurring in blackjack. They plotted the usual calibration (S) curves, but then because OPs based on card frequencies were available, they also looked at mean judgment per OP value (O curve) for each of the three groups. Groups differed little in the general appearance of the curves. Unremarked on by the authors, however, was that the O and S curves lead to opposite conclusions regarding relative degrees of confidence.

Overconfidence and Conservatism as Statistical Phenomena

We can write two empirical relations that summarize all the preceding analyses:

$$E(\mathrm{SP} \mid a < \mathrm{OP} < b) = .5w_s + c_s(1 - w_s) \quad (1)$$

and

$$E(\mathrm{OP} \mid a < \mathrm{SP} < b) = .5w_o + c_o(1 - w_o), \quad (2)$$

where c_s and c_o are the means of the observations in the respective (a, b) intervals, and w_s and w_o $(0 < w_s, w_o < 1)$ are weights reflecting where between .5 and c the respective expected value falls. Except for the extreme intervals, a and b are consecutive multiples of .05 that are not multiples of .1 (so that the midpoints of these intervals are .1, .2, . . . , .9).[4] In words, when OP is constrained to the interval (a, b) the mean of the corresponding SP is between .5 and the mean of that (a, b) interval, and, simultaneously, when SP is identically constrained, the corresponding mean OP is similarly affected. When $a = b$, Equations 1 and 2 represent what is commonly taken as regression to the mean. When, as in the within-subject analyses of the Erev and Wallsten (1993) data, $a = .5$ and $b = 1$ or $a = 0$ and $b = .5$, Equations 1 and 2 represent what Samuels (1991) called reversion to the mean. Note that Equations 1 and 2 cannot simultaneously be true if the correlation between SP and OP is perfect (i.e., 1). Thus, on the very reasonable assumption that judgments have an error component associated with them, the possibility exists that the phenomena of over- and underconfidence are often or primarily statistical consequences of how the data have been analyzed.

We are not arguing that overconfidence and conservatism are necessarily or entirely statistical artifacts. That position would be foolish given that their magnitudes can be systematically varied and that for either type of analysis the effects sometimes invert. We are arguing, however, that the relation between SP and OP in a particular context needs to be established after controlling for random factors in judgment or response. It is possible that in the absence of error only Equation 1 would hold, suggesting real underconfidence, or only Equation 2 would hold, suggesting actual overconfidence. However, other patterns may obtain as well. A pleasant outcome under such circumstances would be that neither Equation 1 nor 2 hold, but that instead true judgments are accurate.

Empirically, the problem is to assess the functional relations between SP and OP as error variance decreases or is controlled. To gain insight into what might be expected, we develop a very general model in the next section and illustrate it with a particular exemplar that incorporates error components of different magnitudes. By investigating its behavior under various contexts designed to represent different experimental conditions, we show that even when true judgments are accurate, results mirror those obtained above as well as others commonly found in the literature. As would be anticipated, the S curves and O curves become more similar to each other as error variance decreases.

Stochastic Models of Probability Estimation

Making use of ideas that date back to Thurstone (1927), the observed effects are easily understood simply by assuming stochastic components in the judgment and response processes underlying SP estimation. We need only three constructs: true

[4] The two extreme intervals are (0, .05) and (.95, 1.0).

judgment, an error distribution, and a response rule. The subject's true judgment of the likelihood of event i, t_i, is the estimate from 0 to 1 the subject would provide if he or she could operate in a fully repeatable, error-free manner. When considering the likelihood of event i at time j, the subject experiences a degree of confidence, x_{ij}, which depends on his or her true judgment and an error component, e_j. That is,

$$x_{ij} = f(t_i, e_j). \tag{3}$$

A monotonic response rule translates this covert feeling, x_{ij}, into an overt response, y_{ij}, in the [0, 1] interval. That is,

$$y_{ij} = g(x_{ij}). \tag{4}$$

Equations 3 and 4 constitute a very broad class of models. Various special cases can be derived by making assumptions about the nature of f, g, and the error distribution. We present and justify a single example and demonstrate that it predicts the observed results.

This model assumes that when considering the likelihood of event i, a subject experiences a degree of confidence in its truth or falsity that is better represented on an unbounded than a bounded scale. That is, $-\infty < x_{ij} < \infty$. Thus, it is reasonable to specialize Equation 3 as an additive combination of error plus a log-odds transformation of t_i:

$$x_{ij} = \ln\left(\frac{t_i}{1 - t_i}\right) + e_j, \tag{5}$$

$0 < t_i < 1$. Error is assumed to be independently, identically, and normally distributed with $E(e) = 0$, and $\mathrm{Var}(e) = \sigma^2$.

The response rule, Equation 4, should be specialized in a manner consistent with the common observation that subjects tend to respond in units that are multiples of .05 or .10 (Budescu, Weinberg, & Wallsten, 1988; Wallsten, Budescu, & Zwick, 1993). One way to accomplish that goal is to assume that x is mapped into the [0, 1] interval by an operation akin to the inverse of Equation 5 and that the nearest rounded value is then provided as a response. Formally, assume that the subject has decided to use the discrete responses $r_0, r_1, \ldots, r_n$ and transforms x_{ij} to y_{ij} by

$$y_{ij} = \frac{e^{x_{ij}}}{1 + e^{x_{ij}}}. \tag{6}$$

Assume further that the subject selects equally spaced response cut-offs, π_k, $k = 1, \ldots, n$, such that $0 < \pi_1 < \ldots < \pi_n < 1$, and chooses a response, r_k, according to the rule,

$$\begin{aligned} &r_0 \text{ if and only if (iff) } y_{ij} \le \pi_1 \\ &r_i \text{ iff } \pi_i < y_{ij} \le \pi_{i+1} \\ &r_k \text{ iff } \pi_k < y_{ij}. \end{aligned} \tag{7}$$

To illustrate the model's predictions, we assume 11 response categories, with probability estimates of $r_k = .025, .10, .20, \ldots, .90$, or .975,[5] and response thresholds at $\pi_1 = .05$, $\pi_2 = .15, \ldots, \pi_{10} = .95$. We assume, finally, that the subject endeavors to use the response categories accurately, or more precisely that the true judgments are always correct. This assumption is implemented by allowing "true" environmental probabilities, p_i, to take on values only of .025, .10, .20, . . . , .90, or .975; and always setting $t_i = p_i$.

In a calibration study, one might imagine these assumptions to describe a subject who places each event into 1 of 11 categories, depending on x, his or her degree of confidence in its truth, and who in the absence of error would operate such that p_i is the proportion of items in category i that in fact are true. In a study in which the OPs are determined by relative frequency considerations or by Bayesian calculations one can imagine the p_i to be the values allowed by the design of the study. Thus, the assumptions involved to implement this model can be considered appropriate for either the calibration or revision-of-opinion paradigm.

To simulate a wide range of subject differences and experimental conditions, we derive predictions from this model for four error variances and three distributions of environmental probabilities. The predictions are in the form of a joint probability distribution over (p_i, r_k), $i, k = 0, \ldots, 10$, which is then summarized in terms of either row or column expected values, depending on which analysis is being mimicked.[6] Calculation of $S(p_i) = E(r|p_i)$ for all i corresponds to analyses yielding the O curves in the previous figures, that is, analyses of SP conditional on OP. Also, calculation of $O(r_k) = E(p|r_k)$ for all k corresponds to analyses yielding the S curves, that is, analyses of OP given SP.

We calculated $S(p_i)$ and $O(r_k)$ for all combinations of three distributions of true probabilities and four levels of error variability across 11 intervals. The three distributions are uniform, U shaped, and W shaped. The uniform distribution represents paradigms, such as Erev and Wallsten's (1993) video game study, in which event probabilities are inferred from observation and a priori are all equally likely. The U-shaped distribution assumes that most items are very probably true (fall in the last interval) or very probably false (fall in the first interval) to represent the situation in most revision-of-opinion paradigms following a reasonable sampling of data. Specifically, for this distribution we assumed that each of the end probability values has a prior probability of 1/3 and each of the nine middle values has a prior probability of 1/27. Finally, to simulate the case of most calibration studies, the W-shaped distribution has large concentrations in the middle (centered at .5) and the end categories to represent a preponderance of items that are known to be true or to be false or for which there is complete uncertainty. On the basis of the data reported by Wallsten et al. (1993), we assigned prior probabilities of .15 to each end value, .30 to the central value, and .05 to the remaining eight values. The four values of σ used were 0.5, 1.0, 1.5, and 2.0.

The resulting S and O curves for the 12 cases are presented in Figure 5, where the simultaneous underconfidence of the O curve and overconfidence of the S curve can be seen in all cases. Note first that their magnitudes depend on the error variance. Underlying true judgments are always accurate, but observed

[5] The natural response values, as evidenced by many studies (e.g., Wallsten et al., 1993), are 0, .10, . . . , .90, 1.0. We are using end values of .025 and .975, respectively, to simplify the analysis and presentation. This discrepancy is of no consequence in anything that follows.

[6] Details of the derivations may be obtained by writing to any of the authors.

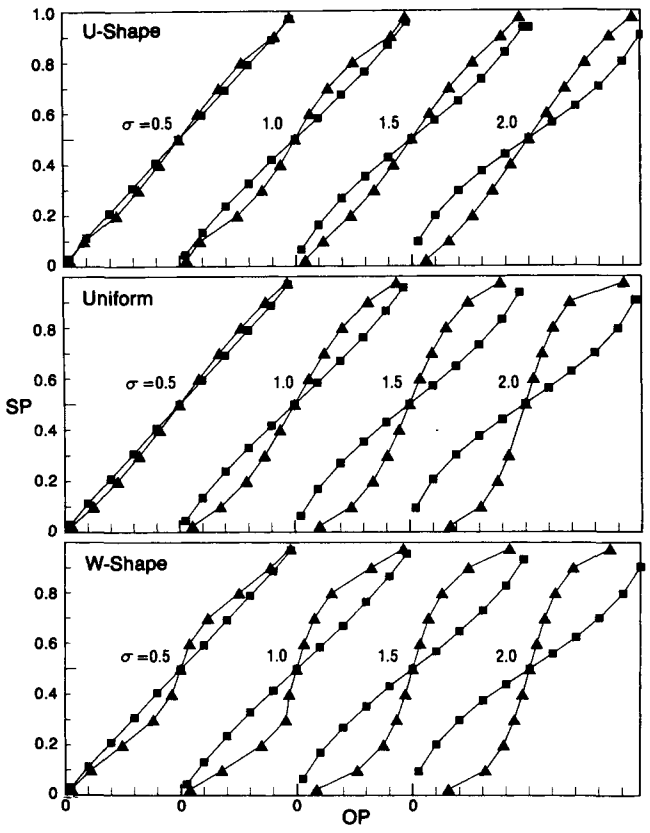

Figure 5. Mean objective probability (OP) as a function of subjective probability (SP; S curve denoted by ▲) and mean SP as a function of OP (O curve denoted by ■) for 12 instances of the log-odds model defined by three prior distributions (U shaped, uniform, and W shaped) and four levels of error variance (σ = 0.5, 1, 1.5, and 2). The abscissa increments in units of .1 beginning at 0, as indicated for each pair of curves.

under- and overconfidence systematically increase as error variance increases. The magnitudes of the effects depend to a lesser degree on the prior distribution of true score values, and moreover there is an asymmetry between the two phenomena. That is, equating for error variance, degree of overconfidence represented by the S curve depends on the distribution of true scores and is smallest in the U-shape case. In contrast, degree of underconfidence represented by the O curve is relatively insensitive to the prior distribution. We verified these observations by calculating average deviations from accuracy, as was done with the $CONF_O$ and $CONF_S$ measures when analyzing the individual data in the Erev and Wallsten (1993) study and present the results in Table 1.

Discussion

We began by pointing to the apparent contradiction in the probability judgment literature between two robust empirical phenomena, overconfidence in most calibration studies and underconfidence (conservatism) in most revision-of-opinion studies. By reanalyzing data from three experiments, we showed that the two results can coexist and that their manifestation may depend on how the data are grouped or conditioned. Judgments suggest underconfidence when they are averaged or presented as a function of independently defined OPs. On the other hand, they appear to reflect overconfidence when objective values are averaged conditional on SPs. Consequently, in many instances, the two effects may simply be artifacts of the methods of data analysis or at the very least be exaggerated by them.

Next, we suggested that these effects are consistent with the notion, hardly new, that judgments, responses, or both are perturbed by error. We illustrated the possibility by means of the log-odds model, as a representative of the class of models expressed by Equations 3 and 4. Particularly striking is the fact that a subject whose true judgments are accurate can appear to be under- or overconfident simply as a function of the magnitude of the error with which those judgments are expressed.

We make no special claims regarding the log-odds model, but we do emphasize its important features: (a) It distinguishes overt response from covert judgment, and (b) it combines true judgment with error in a manner such that Equations 1 and 2 are simultaneously satisfied. We believe that any model meeting conditions (a) and (b) will be a special case of the general model in Equations 3 and 4 and will predict the same overall pattern of results. As general as Equations 3 and 4 are, they represent a broad class of psychological theories, including some proposed elsewhere (e.g., Björkman, in press; Björkman, Juslin, & Winman, 1993; Soll, 1993; Wallsten & González-Vallejo, 1994). Specific assumptions about the nature of the judgment and response process will yield specific instantiations of these equations, which will satisfy conditions (a) and (b) but differ from each other on other predictions. Indeed, we have developed other models from this perspective and will report the results in a subsequent article.

In the remainder of this discussion, we consider the importance of explicitly incorporating error in theories of judgment, paying special attention to the related issues of which of SP and OP should be considered the independent and dependent variables and what are meaningful ways to ask questions about over- and underconfidence. In so doing, we finally help our confused businessman decide whose advice to follow.

Table 1
Mean Relative Confidence Measures ($CONF_o$ and $CONF_s$) for 12 Illustrative Examples of the Log-Odds Model With 11 Response Categories

Prior distribution	σ of error variance	$CONF_o$	$CONF_s$
U shaped	0.5	−0.007	0.004
	1.0	−0.024	0.011
	1.5	−0.045	0.021
	2.0	−0.074	0.043
Uniform	0.5	−0.009	0.020
	1.0	−0.028	0.062
	1.5	−0.052	0.110
	2.0	−0.078	0.155
W shaped	0.5	−0.008	0.037
	1.0	−0.025	0.077
	1.5	−0.050	0.116
	2.0	−0.075	0.153

In virtually every area of psychology, on the basis of the assumption that an organism's behavior depends on the stimulus situation, response measures are analyzed as a function of stimulus parameters. It is a historical curiosity, dating back at least to Adams and Adams (1961), that the opposite has come to be the norm in judgment research. That is, judgment investigators generally look at stimulus characteristics as a function of the response, without apparently recognizing the force of this change in perspective. Authors tend to move between the two approaches, seemingly as a matter of simple convenience (see Griffin & Tversky, 1992; Koriat, 1993; Wagenaar & Keren, 1985, as a few examples). In retrospect, it is easy to identify at least two reasons why this change occurred. First, measures of OP are not available in most interesting judgment situations, and it has become increasingly apparent over the years that generalizations from paradigms in which such measures can be defined to more interesting contexts in which they cannot are tenuous at best. Second, real-world applications of judgment research raise questions about the accuracy of SP estimates, which focus attention naturally on proportion correct as a function of the estimates.

In fact, both methods of analysis are useful. Interpreted properly, they are not contradictory, but they certainly are not equivalent and interchangeable. This point is best understood by reconsidering our earlier data and model analyses, which began with the joint distribution over (p_i, r_k). Mean SP conditional on OP, $S(p) = E(r|p)$, was estimated from the row means of the matrix, and mean OP conditional on SP, $O(r) = E(p|r)$, was estimated from the column means. It is easy to see that these two ways of summarizing the matrix illuminate different of its characteristics, as represented by Equations 1 and 2, respectively, and depending on the question being asked one representation may be more useful than the other. Until now researchers have chosen how to summarize their data primarily as a matter of empirical convenience without regard to theoretical implications. Unless response and judgment error can be greatly minimized or eliminated, that practice is unsatisfactory and can lead to potentially erroneous or misleading conclusions.

Murphy and Winkler (1987, 1992; see also Yates, 1982) made the same point in developing a framework for forecast verification, although on somewhat different grounds. They urged that analyses begin with the joint distribution over (z_i, r_k), where $z_i = 0$ (if the event does not occur) and 1 (if it does) and showed the differences between $E(r|z_i)$ and $E(z|r_k)$. Relating certain of their developments to our terms, let T and F refer to true and false events, respectively. We can think of $E(p|r)$ as the probability of a true event, T, conditional on judgment r. That is,

$$E(p|r) = P(T|r),$$

which from Bayes rule can be written as

$$P(T|r) = \frac{P(r|T)P(T)}{P(r)}. \quad (8)$$

In situations of the sort we have been considering, $P(T)$, the unconditional probability of the event being true, depends on task characteristics only. For example, $P(T)$ will depend on the task or the context when the subject must give a probability judgment regarding the truth of a particular statement, the occurrence of rain, or a specified cause as the source of a set of observations. In contrast, $P(T)$ is not independent of the subject whenever the event being judged depends in part or entirely on a prior choice by the subject. That situation occurs most commonly when the respondent first selects an answer to a question and then gives a probability estimate. For ease of analysis, it seems preferable to work in contexts in which $P(T)$ reflects the task, environment, or knowledge domain only; we restrict our remarks to such contexts only.

In contrast to $P(T)$, $P(r|T)$ and $P(r|F)$, the distributions of the subject's estimates given true and false statements, respectively, are under the subject's control. They reflect his or her judgment and response processes. Thus, the remaining term on the right-hand side of Equation 8, $P(r) = P(r|T)P(T) + P(r|F)[1 - P(T)]$, depends on both the subject and the task. Consequently, without a careful understanding of both the task environment as represented by $P(T)$ and the subject as represented by $P(r|T)$ and $P(r|F)$, one cannot be certain whether $P(T|r)$, or $E(p|r)$, is greater than, less then, or equal to r. Yet this is precisely the question asked for most decision analyses.

Thus, returning to the businessman whose plight illustrated the apparent paradox, it is clear that he is interested in $P(T|r)$, estimated by relative frequency correct conditional on a judgment. Assuming that the common generalization is valid in this case, he should not invest. However, for the reasons just discussed as well as others (see, e.g., Griffin & Tversky, 1992; or Gigerenzer et al., 1991), we cannot be certain of its validity.

Empirical estimates of $E(p|r) = P(T|r)$ can also be of interest for investigations of basic cognitive processes, but its components may be of more direct use. Those under the subject's control include $P(r|T)$, $P(r|F)$, and $P(r|p)$, the respective distributions of r conditional on true and false statements or conditional on OPs when the latter can be independently defined. Therefore, we suggest that these are the measures of most direct interest for basic investigations into the cognitive processes of judgment. That is, the experimenter should control the distributions of true and false statements or of OPs and observe the resulting distributions of responses. This, of course, is consistent with standard practice in virtually all of psychology and follows from the assumption that an organism's behavior depends on the stimulus situation. It is precisely what was done in the studies of probability revision that were essentially abandoned some 15 years ago.

Thus, to summarize, the correct way to look at judgment data depends on the question being asked. Conditionalize on the event state (or on probability when it can be independently defined) when the research is aimed at understanding the underlying cognitive processes themselves. Conditionalize on response when the focus is on accuracy. In either case, the performance measure will not be fully understood without incorporating notions of error.

It is time for basic judgment research to orient away from questions of response accuracy. Rather, theoretically interesting questions within the framework proposed here concern how overt estimates depend on covert judgment and error, how such judgment arises, what factors affect it, and the extent and locus of errors. Models of the sort proposed here are needed to define the constructs and to provide frameworks for drawing inferences about them. From this perspective, theoretical questions

of under- or overconfidence should be refocused. That is, controlling for response error, judgment error, or both, we can ask what conditions cause underlying judgment to be accurate or to be more or less extreme than warranted by the available information? To date, we do not know the answer.

References

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological Review, 68,* 33–45.

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

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

Budescu, D. V., & Wallsten, T. S. (1987). Subjective estimation of vague and precise uncertainties. In G. Wright & P. Ayton (Eds.), *Judgment forecasting* (pp. 63–82). New York: Wiley.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 281–294.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representations of human judgment* (pp. 17–52). New York: Wiley.

Erev, I., Bornstein, G., & Wallsten, T. S. (1993). The negative effect of probability assessment on decision quality. *Organizational Behavior and Human Decision Processes, 55,* 78–94.

Erev, I., & Wallsten, T. S. (1993). The effect of explicit probabilities on decision weights and on the reflection effect. *The Journal of Behavioral Decision Making, 6,* 221–241.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90,* 239–260.

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

Gigerenzer, G., Hoffrage, U., & Kleinbölting, 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.

Howell, W. C. (1972). Compounding uncertainty from internal sources. *Journal of Experimental Psychology, 95,* 6–13.

Howell, W. C., & Burnett, S. A. (1978). Uncertainty measurement: A cognitive taxonomy. *Organizational Behavior and Human Performance, 22,* 45–68.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11,* 143–157.

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

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609–639.

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). Cambridge, England: Cambridge University Press.

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review, 115,* 1330–1338.

Murphy, A. H., & Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting, 7,* 435–455.

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

Pitz, G., Downing, L., & Reinhold, H. (1967). Sequential effects in the revision of subjective probabilities. *Canadian Journal of Psychology, 21,* 381–393.

Rapoport, A., & Wallsten, T. S. (1972). Individual decision behavior. *Annual Review of Psychology, 23,* 131–176.

Rapoport, A., Wallsten, T. S., Erev, I., & Cohen, B. L. (1990). Revision of opinion with verbally and numerically expressed uncertainties. *Acta Psychologica, 74,* 61–79.

Ronis, D. L., & Yates, 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.

Samuels, M. L. (1991). Statistical reversion toward the mean: More universal than regression toward the mean. *The American Statistician, 45,* 344–346.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6,* 649–743.

Soll, J. B. (1993). *Determinants of miscalibration and over/underconfidence: The interaction between random noise and the ecology.* Unpublished manuscript.

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

Wagenaar, W. A., & Keren, G. B. (1985). Calibration of probability assessments by professional blackjack dealers, statistical experts, and lay people. *Organizational Behavior and Human Decision Processes, 36,* 406–416.

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

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probabilistic judgments. *Management Science, 39,* 176–190.

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

Winkler, R. L., & Murphy, A. (1973). Experiments in the laboratory and the real world. *Organizational Behavior and Human Performance, 10,* 252–270.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Decision Processes, 30,* 132–156.

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

Received July 12, 1993
Revision received November 4, 1993
Accepted December 6, 1993 ■