

Evaluating and Combining Subjective Probability Estimates

THOMAS S. WALLSTEN,¹ DAVID V. BUDESCU,² IDO EREV³
AND ADELE DIEDERICH⁴

¹*University of North Carolina, USA*

²*University of Illinois, USA*

³*The Technion, Israel*

⁴*Universität Oldenburg, Germany*

ABSTRACT

This paper concerns the evaluation and combination of subjective probability estimates for categorical events. We argue that the appropriate criterion for evaluating individual and combined estimates depends on the type of uncertainty the decision maker seeks to represent, which in turn depends on his or her model of the event space. Decision makers require accurate estimates in the presence of aleatory uncertainty about exchangeable events, diagnostic estimates given epistemic uncertainty about unique events, and some combination of the two when the events are not necessarily unique, but the best equivalence class definition for exchangeable events is not apparent. Following a brief review of the mathematical and empirical literature on combining judgments, we present an approach to the topic that derives from (1) a weak cognitive model of the individual that assumes subjective estimates are a function of underlying judgment perturbed by random error and (2) a classification of judgment contexts in terms of the underlying information structure. In support of our developments, we present new analyses of two sets of subjective probability estimates, one of exchangeable and the other of unique events. As predicted, mean estimates were more accurate than the individual values in the first case and more diagnostic in the second. © 1997 by John Wiley & Sons, Ltd.

Journal of Behavioral Decision Making, 10, 243–268 (1997)

No. of Figures: 2 No. of Tables: 3 No. of References: 74

KEY WORDS subjective probability; combining judgments; exchangeability; uncertainty

INTRODUCTION

For many real-world decisions, ranging from the mundane to the crucially important, it is necessary to combine information from multiple sources prior to taking action. The information may be heavily

Correspondence to: Thomas S. Wallsten, Department of Psychology, University of North Carolina, Chapel Hill, NC 27599-3270, USA. E-mail: tom.wallsten@unc.edu

statistical, as with the outcomes of experimental or epidemiological studies, or it may be in the form of expert interpretation of indicators, as regarding the likelihood of earthquakes, severe weather conditions, or the outcome of a particular medical procedure on a 68-year-old man with a history of coronary disease. The issues regarding how to combine information from multiple and disparate sources are so important and so difficult, that a special committee of the US National Research Council studied them and issued a report with recommendations regarding procedures and future research on the topic (National Research Council, 1992).

In particular (references to follow), researchers in management science, operations research, and statistics have sought to develop formal procedures based on normative considerations for combining expert forecasts or judgments. Behavioral scientists (primarily, but not exclusively, psychologists) have empirically compared the various methods to each other as well as to individual judgments.¹ Our present research seeks to unite the best of the behavioral and the normative approaches by providing a cognitive foundation for combining multiple judgments. Specifically, our goals are to:

- (1) Develop methods for combining subjective probability judgments that are based on sound theoretical principles, are easily operationalized, and yield results that are of maximal diagnostic value.
- (2) Ground the principles and methods in cognitive models that describe how individual judges form and express their opinions, and that are sensitive both to judges' levels of knowledge and to the nature of the available information.
- (3) Empirically test the underlying cognitive models where necessary.
- (4) Empirically evaluate our combination rule (or rules) with behavioral data and computer sensitivity runs, as appropriate.

This paper represents a step toward meeting those goals.

We consider here only the problem of combining subjective probability estimates for categorical events. A proper solution to this aspect of the more general combination problem has the potential to both improve the quality of important decisions and increase our understanding of basic human judgment processes. We begin by providing examples of situations in which such judgments must be combined. Then we discuss criteria for assessing the quality of probability estimates and provide a brief review of the relevant theoretical and empirical literature. Next, we develop our theoretical approach and present empirical results in support of it. Finally, we discuss the implications of our developments, as well as some of the remaining open problems.

Categorical events have a finite number of possible outcomes. By focusing on a particular outcome, O , and its complement, $\text{not-}O$, we define binary events, which can be termed generically as true or false, or as occurring versus not occurring. Examples include a statement being true or false, a particular patient surviving or succumbing during surgery, or two countries settling a particular dispute within a particular time frame. In all these cases and many others the event is well defined, but the uncertainty regarding it cannot be represented by statistical or relative frequency information. Rather, the uncertainty must be obtained and expressed as a matter of expert judgment, and in many (probably most) cases experts will disagree.

Consider the medical example. A very sick individual may be willing to undergo a certain surgical procedure if his survival chances are sufficiently great, say at least 80%. He consults two specialists. On the basis of the same medical information, one expert estimates the chances as 60% and the other as

¹ Social psychologists have contrasted judgments reached independently to those reached via interacting groups, but that work does not relate to our current concerns. In his excellent review of that literature, Locke (1987) concluded among other things that group interaction is useful in many ways, but does not substitute for ultimately combining separate judgments by means of a formal rule.

95%. Their estimates may differ because the experts have different views of the patient's condition, of the surgical staff's competence, or of the relevant medical literature. The patient may discuss these matters with the two experts, but ultimately the issues are beyond his range of knowledge and he is left with their two probability estimates. How should he combine them? Should he take an arithmetic mean, perhaps weighting each judgment to reflect perceived or measured competence? Would a (weighted) geometric mean of the odds be better? Or would yet some other combination procedure be optimal? One can readily see the same problems occurring in many other contexts.

Our formulation of the combination problem rests on the claim that rules for aggregating subjective probability estimates should be sensitive to both the structure of the information base supporting the estimates and the cognitive processes of the judges who are providing them. The former point is well understood and often discussed (e.g. Chapters 11–13 of Cooke, 1991; French, 1986; Genest and Zidek, 1986). The latter one seems obvious and is implicit in much of the literature, yet to our knowledge, it has not been made explicitly before. To develop the points simultaneously, we next discuss criteria for assessing the quality of individual or aggregated subjective probability estimates. With that background, we then briefly consider the mathematical literature on combining expert judgment, the behavioral literature comparing combined to individual judgments, and the cognitive literature on the structure of individual judgments. All this material sets the stage for the development of our approach in the following section.

Criteria of quality²

In the case of physical dimensions, the quality of a set of subjective estimates depends primarily on their accuracy, which assuming agreement on the issue of scaling is straightforward to define and measure. For example, on an ordinal scale, subjective estimates are accurate when their ranking coincides with that of an independent physical measure. Deviations from accuracy are scored by an index of rank inversions. On a particular ratio scale (e.g. grams of mass), subjective estimates are accurate when they equal the physical measure. Deviations from accuracy are scored by an index of closeness, such as the root-mean-squared deviation between the subjective and objective measures, perhaps normalized by the standard deviation of the physical measures.

The issue of quality is considerably more complex, however, when it is probability rather than a physical characteristic that is being estimated. As De Finetti (1974) and others have argued, probability is not a property of the external world, but the representation of a coherent decision maker's opinions about a set of events. We suggest that the appropriate criterion of quality varies with the type of uncertainty the decision maker seeks to represent. Heath and Tversky (1991) distinguished two types of uncertainty: aleatory (associated with external chance factors) and epistemic (associated with internal lack of knowledge). Others have made similar distinctions (e.g. Budescu and Wallsten, 1987; Howell, 1971; Kahneman and Tversky, 1981; Vesely and Rasmuson, 1984; Wallsten, 1990; Whitfield and Wallsten, 1989). This neat division, however, is not absolute, as it depends on the decision maker's model of the events in question. Moreover, when he or she is unsure about the model to apply, or considers more than one model plausible, then the uncertainty is both aleatory and epistemic in some combination.

Uncertainty is aleatory when the decision maker can partition observations into equivalence classes of exchangeable events. The notion of *exchangeability* was developed by De Finetti (1937/1964) and is described lucidly by Cooke (1991, Chapter 9). Briefly, a sequence of binary events (coding each

² The ideas in this section have benefited greatly from comments from and discussions with Bob Clemen, Craig Fox, Claudia González-Vellajo, Michael Lavine, In Jae Myung, and Jack Soll, as well as from the very thoughtful reactions of the audience to a presentation by the first author at a Duke Fuqua School of Business Decision Analysis Workshop.

outcome as 0 or 1) is exchangeable (or the sequence consists of exchangeable events) if for any subset length n of the sequence containing r 0s and $n-r$ 1s, the decision maker considers all permutations of 0s and 1s to be equally likely.

Note that two people may disagree on whether a set of events is exchangeable. Thus, a person who knows nothing about the seasonal weather pattern in a particular location may consider all permutations of 50 rainy and 50 non-rainy days to be equally likely for a given 100-day period, while someone with knowledge of the area may consider some permutations to be more likely than others. The sequence is exchangeable for the first person, but not for the second. More generally, decision makers may define equivalence classes of exchangeable events with differing levels of precision, depending on their knowledge and goals. Thus, mortality tables may group people by country of residence only; country and gender; country, gender, and age; country, gender, age, and smoking history; etc. Any uncertainty about the partition to use is epistemic, but given a particular partition, the uncertainty within it is aleatory. The event probabilities are expected to be increasingly distinct as the equivalence classes become more precisely defined. Only when there is universal agreement on the definition of the equivalence classes might we say that the uncertainty is purely aleatory.³

Events that the decision maker does not consider exchangeable with other events are unique. Thus, most people will consider the event that Amsterdam is north of Paris to be unique. Any uncertainty they have regarding its truth is epistemic. However, for individuals knowing nothing about the geography of Europe, this event may be exchangeable with many others concerning the relative latitudes of pairs of European cities. The uncertainty for them is aleatory.⁴

Returning to the question of quality, we assume that decision makers want probability estimates that (1) are coherent and (2) yield maximum expected value or expected utility decisions (Clemen and Murphy, 1990). Consistent with those requirements, we suggest that decision makers want accurate probability estimates when they treat observations as belonging to equivalence classes of exchangeable events and diagnostic estimates when they treat the observations as unique. When decision makers are unsure of how to define the equivalence classes, or believe that others may define them in ways that can lead to improved resolution, then they want estimates that are maximally diagnostic (highly resolved, in the language associated with Brier score partitions — Yates, 1982, 1994) while still being accurate.

Generally speaking, estimates are accurate if they match the (expected) relative frequencies of the observations in each equivalence class. They are diagnostic if they allow one to distinguish events that are true from those that are false. Of the many possible indices of accuracy and diagnosticity, we propose ones that accord with common practice.

A natural index for accuracy is the mean squared deviation over equivalence classes between the estimated probabilities and the event relative frequencies. Specifically, let $a_1, a_2, \dots, a_i, \dots, a_I$ be I distinct events (equivalence classes of observations) and $p(a_i)$ the (expected) relative frequency of a_i . Further, assume that a judge estimates the probability of a_i on each of N_i occasions, and let $f(a_i)$ denote a central statistic (e.g. mean, median, or mode) for those N_i estimates. Finally, let $N = \sum_{i=1}^I N_i$ be the total number of estimates. Then, the accuracy index is

$$CI' = \frac{1}{N} \sum_{i=1}^I N_i (f(a_i) - p(a_i))^2 \quad (1)$$

³ Universal agreement may be hard or impossible to come by, as Davidson, Suppes, and Siegel (1957) discovered when seeking a fair die that all research participants treated as fair. But we can imagine the concept in principle.

⁴ One might argue that because Amsterdam either *is* or *is not* north of Paris, any uncertainty about that event necessarily is epistemic and not aleatory. By the same token, one then would be forced to argue that if a coin had been tossed and *did* or *did not* land heads, uncertainty about that event is necessarily epistemic. The issue is complex and further discussion of it is beyond our scope.

We use the notation CI' to emphasize the similarity of this index to the calibration index, CI , obtained from Murphy's (1973) decomposition of the quadratic scoring rule, which we will discuss shortly. Generally, one would set $f(a_i)$ equal to the mean of the estimates of a_i but depending on one's model of the estimation process, an alternative central measure might be preferable.

Turning now to an index of diagnosticity, we propose one closely related to Yates' (1982) slope, which he defined in conjunction with his covariance decomposition of the quadratic scoring rule. Yates' slope is the difference between the means of the estimated probabilities conditional on true and false events, respectively, and serves as an index of the degree to which subjective probabilities near 1 are assigned to true events and near 0 are assigned to false events. In addition, Yates (1982) emphasized, and Yates and Curley (1985) demonstrated, that the most useful estimates are those that maximize the slope while minimizing the conditional variances of the judgments. We propose combining these two desiderata into one index of diagnosticity, DI' . Recall that for this index, we are combining observations not into equivalence classes of exchangeable events but into two classes according to whether they are true or false (occur or do not occur). Let \bar{f}_T and \bar{f}_F be the mean probabilities assigned to true and false events, respectively, and let s_f^2 be the pooled variance of the two conditional distributions. Then,

$$DI' = \frac{\bar{f}_T - \bar{f}_F}{s_f} \quad (2)$$

Note that DI' is similar in structure to d' from signal detection theory, in that it indexes the difference between two means relative to a common standard deviation. Moreover, it does not have the problems that Yaniv, Yates and Smith (1991) attribute to DI , the discrimination index from Murphy's (1973) partition of the quadratic rule.⁵

Finally, we need a quality index for situations in which the decision maker believes the events are not truly unique, but is unsure of the best equivalence classes to use. The most widely used index is the quadratic loss function, often called the Brier probability score (PS), to which we have already referred. PS is calculated on the individual estimates without regard to an a priori equivalence classification. Letting N be the number of observations for which probabilities are estimated, f_n the estimate assigned to the n th observation and d_n an outcome index (equal to 1 if event n occurs and 0 if it does not), for $n = 1, \dots, N$,

$$\overline{PS} = \frac{1}{N} \sum_{n=1}^N (f_n - d_n)^2 \quad (3)$$

Thus, \overline{PS} is the mean squared deviation between the estimates and the indicator variables.

Murphy (1973) has decomposed equation (3) into components that reflect overall forecasting difficulty, calibration, and discrimination by classifying events according to their assigned probability estimates. It has become common practice in judgment research to classify events in this way, and that is a reasonable strategy when one has no other basis for classifying them. But the approach may lead to

⁵ The problems are that DI is limited by the proportion of true events in the sample being judged and biased as a function of the number of judgment categories used. Equation (2) is nevertheless not a perfect index. For one thing, it could prove problematic if the variances of the estimates conditional on true and false events, respectively, are substantially different from each other, in which case, s_f^2 might be misleading. Second, DI' can be deceiving in cases where there is no overlap in the subjective probabilities accorded true and false events. The lack of any overlap means that the estimates are perfectly diagnostic of whether events are true or false. Yet two sets of estimates may share this characteristic but have distinct DI' values. An index that provides a measure of overlap might handle both problems and in the end be more useful than equation (2), but we defer that line of development here.

different conclusions than when one determines the equivalence classes on other grounds. This point was emphasized by Erev, Wallsten and Budescu (1994), further developed by Wallsten (1996), and further illustrated by Budescu, Erev and Wallsten (1997). It is also apparent in the work of Yates (1982) and Murphy and Winkler (1992). Murphy's decomposition is widely used (see Yates, 1982, 1994), and we present it here to compare its components with CI' and DI' .

Using Yates' (1994) notation, equation (3) can be rewritten as

$$\overline{PS} = \bar{d}(1 - \bar{d}) + CI - DI \quad (4)$$

where \bar{d} is the fraction of events that are true or occurred and, therefore, $\bar{d}(1 - \bar{d})$, the variance of the outcome index, is a measure of forecasting difficulty. CI and DI are the calibration index and discrimination index, respectively, which assuming the use of discrete probability judgment categories, are:

$$CI = \frac{1}{N} \sum_{j=1}^J N_j (f_j - \bar{d}_j)^2 \quad (5)$$

and

$$DI = \frac{1}{N} \sum_{j=1}^J N_j (\bar{d}_j - \bar{d})^2 \quad (6)$$

In these equations, J is the number of probability estimation categories, f_j now is the category value, N_j is the number of events the judge places in category j , and \bar{d}_j is the proportion of those events that were true or occurred.

Note that, in general, $CI \neq CI'$, the former being conditioned on probability estimation categories and the latter on independently defined event equivalence classes. However, when CI' is calculated by setting $f(a_i)$ equal to the mean of the estimates of a_i (see equation (1)), then $CI = CI'$ under certain conditions. These conditions are that the number of estimation categories equals the number of equivalence classes, the category labels match the equivalence class long-run relative frequencies, and the association matrix denoting the number of times each label is assigned to each class is symmetric about the diagonal.⁶ The closer this matrix is to symmetric, the closer will be the two indices. We conjecture, but have not proved, that, in general, when pooling multiple sets of probability estimates, procedures that minimize CI' also will minimize CI , and conversely.

It is also true that, in general, $DI \neq DI'$. In fact, the relation between DI' and DI is indeterminate because the former is based on the particular subjective probability categories that are used (the f_j) and the latter is based on the proportion true in each category (the \bar{d}_j). Again, we suspect, but have not proved, that rules for combining probability estimates will have equivalent effects on DI' and DI .

To summarize, we suggest that the appropriate quality measure for individual or aggregated subjective probability estimates depends on the decision maker's model of the events being estimated. CI' is a suitable index when the decision maker is comfortable with a particular partitioning of the event space into equivalence classes. DI' will do when the decision maker treats the events as unique. And \overline{PS} is suitable when the decision maker neither considers the events unique nor has a strong

⁶ To see that $CI = CI'$ under the stated conditions, consider a square $F \times P$ data matrix with rows, $F = (f_1, f_2, \dots, f_J)$, denoting subjective estimates and columns, $P = (p_1, p_2, \dots, p_J)$, denoting equivalence class long-run relative frequencies, such that $f_1 = p_1, f_2 = p_2, \dots, f_J = p_J$. Let n_{ij} index the number of times estimate f_i is associated with the equivalence class p_j . Now write equations (1) and (5) for CI' and CI , respectively, in terms of this data matrix. Note that $CI - CI' = 0$ when $n_{ij} = n_{ji}$.

opinion about a particular partitioning of the universe of events. Rules for combining multiple judges' probability estimates should be evaluated according to whether they systematically improve the desired index as the number of judges increase.

We must make a few additional related points before turning to a brief literature review. One is that we are not the first to suggest that criteria other than a global probability score (the Brier or any other proper scoring rule) may be of primary importance. For example, Winkler and Murphy (1968) distinguished normative from substantive goodness of probability assessments, the former being satisfied by estimates that conform to the rules of probability theory and the latter by estimates that predict events in the world. They implicitly gave greater weight to substantive (indexed here by DI or DI') than normative (CI or CI') goodness by pointing out that substantively good assessments can be made normatively good via rescaling, but the opposite cannot be done. Yaniv *et al.* (1991) wrote, 'It can be argued that, given two probability judges who are equally well calibrated [in the sense of equation (5)], the better judge is the one who demonstrates greater discrimination skill' (p. 614). Similarly, Winkler and Poses (1993) wrote 'From a practical standpoint, . . . , the level of discrimination exhibited by probabilities seems more important than their calibration' (p. 1526). What may be new here is the attempt to tie the desired quality index to the decision maker's model of the events rather than to rely more generally on a global probability score. But as Winkler and Murphy (1968) pointed out, alternative equally defensible probability scores do not always lead to the same ranking of probability assessors.

In this regard, as the authors just quoted make clear, there is no need for a decision maker to commit to a particular quality index and ignore the others. For example, a decision maker who desires estimates that minimize CI' may also want to keep track of DI' . If DI' is substantially greater for forecaster A than B, A may be employing a more useful event partitioning, and the decision maker may prefer his or her probability estimates (and equivalence class definitions). Of course, a combination of A's and B's estimates is likely to be better than either one alone. It is to this topic that we now turn.

Mathematical literature on combining judgments

The literature in management science, operations research, and related disciplines considers forecasts much more broadly than the single topic of subjective probability judgments for categorical events, on which we are focusing here. Examples of other types of quantities with which authors are concerned include point estimates of economic indices, future time series, or subjective probability distributions over continuous variables. We limit our brief review only to those aspects of the literature that bear on our problem.

Clemen (1989) and Genest and Zidek (1986) have provided excellent reviews with annotated bibliographies that cover virtually all but the most recent contributions (see also Marley, 1991). Clemen considered probability distributions as one of many types of variables that might be combined, while Genest and Zidek focused on them almost exclusively. Perhaps the most comprehensive treatment of combining probability judgments is by Cooke (1991), who devotes a substantial share of his book to the topic. Genest and Zidek distinguished combination rules based on axiomatic from those based on Bayesian considerations. Cooke distinguished what he called classical from Bayesian approaches to combining expert judgment.

We will consider the axiomatic perspective only very briefly, and then the classical and Bayesian in a little more detail. A point we want to emphasize at the outset is that when applied to expert judgments (as opposed, for example, to the output of statistical models), all the approaches require implicit or explicit assumptions about the judges' knowledge and abilities as well as about the structure of the underlying information base. One such assumption is that judges' levels of expertise can be assessed in

terms of the relations between their probability estimates and the actual pattern of event outcomes. Another is that the dependence among the individual estimates, perhaps due to the judges relying on common information, can be estimated. Our developments derive directly from assumptions about judges and the information base, but from much weaker ones than normally are made.

The axiomatic approach to probability pooling stipulates qualitative conditions that a decision maker may require of any combination rule and then proceeds to derive rules that satisfy these requirements. Examples of such qualitative conditions include that if all experts agree on the probability of an event, then the combined judgment should yield that result (the unanimity principle), and if all experts agree that certain events are independent, then that independence should be preserved in the combined result. While axiomatic methods have provided some insights, commentators generally have been critical. In a series of reactions to an influential paper by Morris (1983), Clemen (1986), French (1986), Lindley (1986), Schervish (1986), and Winkler (1986) raise a variety of concerns about the overall approach. Chief among them are that it does not provide the decision maker any freedom to incorporate his or her own assessment of the judges' abilities or of the level of dependence among the judges. Genest and Zidek (1986) summarize additional concerns in the form of unreasonable necessary implications of the axioms (e.g. some aggregation methods imply that individual judges may become dictatorial). In one way or another, these authors all come down in favor of the Bayesian approach (which, in fact, Morris, 1977, also proposed).

Turning to the Bayesian approach, Bayes' Rule specifies how the probability of a discrete (or the density over a continuous) event, E , should be revised upon the observation of a related event, D . Although the rule itself is uncontroversial, enormous debate has emerged regarding its application when E is a unique event in the real world and D is a datum that aids in predicting it. The controversies concern philosophical and practical issues of defining and measuring probabilities of events for which relative frequencies are not available either in fact or in principle. Nevertheless, many authors claim that the correct way for decision makers to combine multiple probability judgments is to treat the judgments as data regarding the event in question, and then to use Bayes' Rule to revise their own subjective probabilities of the event. According to this approach, decision makers represent their personal assessments of the judges' abilities and interdependencies by means of the conditional probabilities they accord the data (i.e. the experts' judgments). The assessments may depend on the judges' forecasting histories with similar events or they may be based on other factors. Given a suitable history of forecasts, Clemen and Winkler (1987) provide a method for combining conditionally correlated probability estimates based on Lindley's (1982) Bayesian log-odds model for frequency calibration of subjective estimates. More recently, Clemen and Winkler (1993) suggested a flexible modeling approach to Bayesian aggregation of possibly dependent estimates that makes use of influence diagrams. Theoretical and empirical considerations in modeling dependent estimates and establishing the decision maker's conditional likelihoods remain matters of considerable debate (see e.g. the responses by Morris, 1986, or Shafer, 1986, to the Genest and Zidek review, as well as the discussion by Cooke, 1991, Chapters 13 and 15). An interesting point is that under suitable assumptions (e.g. Bunn and Mustafaoglu, 1978), the Bayesian approach prescribes that the combined probability be calculated as a weighted average of the experts' judgments.

Calculating a weighted average of experts' estimates, using weights based on proper scoring rules is what Cooke (1991) calls the classical approach. Just as does the Bayesian, this approach requires subjective input from the decision maker, who must select the scoring rule to use and possibly the context for calculating experts' scores. Scoring rules differ in the extent to which they reward good calibration and high diagnosticity (what Cooke calls low entropy), and different scores may yield different weights. Cooke proposes and justifies a particular weighting scheme. He goes on to suggest that if the experts do not have a history of providing probability estimates regarding the variable or events in question, the decision maker must select the 'seed variables' for experts to estimate so that

prior weights can be established. While this solution may be problematic, it is not any more so than corresponding solutions are with the Bayesian methods. Little work seems to have been done within the framework of weighted averaging rules to handle problems arising from conditionally correlated judgments.

Finally, we mention a new approach for combining judgments based entirely on the principle of maximum entropy (Levy and Deliç, 1994; Myung, Ramamoorti and Bailey, in press). It has emerged only recently, and it is too soon to judge its empirical success.

The empirical literature on combining estimates

As a point of departure, we should note that there is a considerable literature on combining subjective estimates of physical dimensions. Over 70 years ago, Gordon (1924) demonstrated that when subjects rank ordered very closely spaced weights according to their perceived heaviness, the means of the ranks correlated more strongly with the objectively correct ranking than did the judgments of the individual subjects. The means were in fact at least as accurate as the judgments of the best subjects. This basic result has been replicated many times (see Zajonc, 1962, for a review and reanalysis of a number of studies on this topic). Lorge *et al.* (1958) discuss these plus other studies in which individuals gave numerical estimates, rather than rank orderings, of measurable quantities. The findings have been consistent: the means of multiple judgments are more accurate than those of individuals, generally even more accurate than the best of the individual judgments. Similarly, Clemen's (1989) review showed that average forecasts of quantities such as economic indicators outperform the individual forecasts.

It is not hard to imagine that when many individuals estimate perceptual, objectively measurable quantities, they all form the same underlying impression of each stimulus except for an independent and unbiased random component. If that is so, the law of large numbers implies that the mean impression will converge on the correct value as the number of judgments increases. The matter is more complicated if individuals' errors are correlated, perhaps due to common perceptual or judgmental deficiencies. Nevertheless, convergence in the direction of the correct result is still expected on statistical grounds (see Hogarth, 1978). Quite distinct from any model of the error process, McNees (1992) showed that when estimates or forecasts are assessed by their squared deviation from the actual value, means and medians must always appear relatively accurate. These results suggest that average probability estimates will outperform the individual components when the criterion is to minimize CI' (or CI), but provide no guidance on what to expect when the criterion is DI' (or DI) or a probability score such as \overline{PS} . Nor do they suggest what Bayesian aggregation might do.

With this background, we can turn to the evidence comparing the accuracy and performance of individuals' probability judgments to those of group means. Some of the research is reviewed by Clemen (1989) and by Hogarth (1977). We know of no studies that have compared individual and group mean estimates for exchangeable events, although we present such a study below. Our brief review of illustrative studies, therefore, is confined to situations in which judges estimated the probabilities of unique events.

Winkler (1971) had subjects give probability judgments regarding the outcomes of football games. By a variety of probability scores, the means of the judgments outperformed all the individuals. In another study, Clemen and Winkler (1987) compared various averaging and Bayesian pooling procedures for precipitation probability forecasts issued by National Weather Service forecasters. Averages (of simple probabilities, as well as of log-odds) outperformed both the Bayesian methods and the individual estimates.

Goldberg (1970) analyzed data collected by Meehl (1959), in which clinical psychologists rated the likelihood on an 11-point scale that psychiatric patients were psychotic rather than neurotic on the

basis of their MMPI (Minnesota Multiphasic Personality Inventory) profiles. Because the actual diagnoses of the patients were known, the ratings could be validated. Goldberg's main point was that the ratings of individual experts were represented well by linear multiple regression equations. A side point in his article, but of special relevance here, was that the simple arithmetic mean of the ratings correlated better with the outcome variable (i.e. better discriminated psychotic from neurotic patients) than did the individual ratings.

More recently, Curtis, Ferrell, and Hillman (1988) had four radiologists indicate whether they thought an abnormality was or was not present in each of 45 excretory urograms, and assess confidence in their diagnosis on a 3-point scale. Effectively, then, experts provided confidence ratings on a 6-point scale. The authors looked at proportions of correct diagnoses as a function of the number of ratings combined, K , and the method of combination. For all methods, including simple averaging, proportion correct increased with K .

Winkler and Poses (1993) provided a very detailed analysis of the effects of combining expert judgments. They had physicians in an intensive care unit of a teaching hospital estimate patients' probability of survival as soon after each patient's admission as was feasible. For each patient, an intern, a critical care fellow, a critical care attending physician, and the primary attending physician independently provided a survival probability. The authors considered the individual physician estimates ($K = 1$), as well as the mean estimates for all possible combinations of $K = 2, 3$, and 4 physicians. Translating their analyses into our notation, on average, \overline{PS} , CI , and DI' all improved as K increased from 1 to 4, as did DI' , which we calculated from the data they provided.

The Winkler and Poses study also speaks to the issue of conditionally correlated judgments. Their four physician groups probably differed in levels and types of expertise, as well as in the information on which they focused. The interns and critical care fellows were much less experienced overall than the critical care and primary attending physicians, who themselves had very different backgrounds and perspectives. The DI' values for the four groups were 1.60, 1.42, 1.62, and 1.87 for the interns, fellows, critical care attendings, and primary care attendings, respectively. While the mean probability estimates improved on average as K increased from 1 to 4, not all combinations of physicians performed equally well at $K = 2$ or 3. Specifically, although the interns individually were almost as good as the critical care attendings, they contributed the least to any combination of estimates. For $K = 2$, the mean estimates for the two groups of attendings outperformed all other sets of means, while the means for all pairs of groups that included the interns performed the most poorly. Similarly, at $K = 3$, the best groups were those that included the two attendings and the worst were those that included the interns. We can speculate that this pattern occurred because (1) experienced individuals from two different specialty areas are most likely to have different perspectives and focus on distinct indicators, and (2) the interns are most likely to have learned from their mentors and therefore have the greatest overlap in knowledge bases with them.

In an impressive paper, Graham (1996) analyzed the effects of combining up to nine economists' probability estimates of negative economic growth (as measured by the real GNP — gross national product) for the current quarter year, as well as one and two quarters ahead.⁷ He combined the estimates by taking simple means, using Clemen and Winkler's (1987) generalization of Lindley's (1982) Bayesian log-odds pooling procedure for conditionally correlated probabilities, and by a third procedure that he developed. For this BINARY50 method, he dichotomized each expert's probabilities according to whether they were above or below 0.50, assigned a single fitted probability to all the low values and another to all the high ones per forecaster, and then used a version of Bayes' Rule for

⁷ Many more economists participated in the Survey of Professional Economic Forecasters. Graham limited his data set to the nine who provided forecasts for at least 50 of the 91 possible quarters, and not all nine made forecasts on all occasions.

correlated binary conditional probabilities. Using the global \overline{PS} index as well as three other global indices, all the pooled estimates tended to outperform the median forecaster for the current quarter, one period ahead and two periods ahead, but for some indices and some periods this was not true. In terms of CI , the BINARY50 and the average values consistently outperformed the median forecaster, whereas the Bayesian log-odds value did not. With regard to DI , the BINARY50 and average values exceeded the median forecaster for two of the three periods and the log-odds result did for one. Graham provided data that allowed us to calculate DI' for the current quarter and two quarters ahead. This index yielded the same conclusions as did DI .

To summarize, studies of experts in their domains of expertise and of ordinary subjects in laboratory experiments have shown that mean probability estimates (or, where tested, mean log-odds) of unique events outperform individuals' estimates in terms of various criteria. Bayesian methods, despite their possibly stronger philosophical foundation, have not proved notably better than simple averaging. Mean estimates of unique events are more diagnostic (DI and DI' are greater) and calibration tends to be, but is not uniformly, better (CI smaller). McNeese's (1992) arguments regarding the necessary improvement in squared error functions resulting from taking central tendencies might apply to the CI index (depending on how one defines the equivalence classes), but they do not apply to DI or DI' . Nor can one argue that improvement in the latter indices is due simply to averaging out random errors in quantities that are otherwise held in common by all the judges. Underlying judgments vary in such cases, depending on how individuals evaluate the information, what they know of the domain, their theoretical perspective, and on a host of other factors. We propose that the positive effects of averaging probability estimates can be understood, taken further, and built upon for practical applications only by incorporating suitable models of the human judgment process itself.

Individual judgment processes

A thorough summary of this literature would take us too far afield. Instead, we will concentrate only on laying the foundation for our approach to the issue of combining estimates. Readers desiring a more substantial background are referred to a recent thoughtful review by McClelland and Bolger (1994) or a slightly older one by Keren (1991).

To a large extent, current research on individual estimation processes has been guided by the implicit assumption that the diagnosticity of a set of estimates is limited by the judge's level of knowledge, while calibration depends on how he or she translates that knowledge into overt responses. Consequently, the research has focused more strongly on issues of calibration and directions of miscalibration than on issues of diagnosticity. The single finding that has driven most modern research on individual judgment processes is that when relative frequencies true conditional on probability estimates of unique events (\bar{d}_j) are compared to the estimates themselves (f_j), the functions generally increase monotonically, intersect the diagonal in the neighborhood of the midpoint, and have slope less than 1, i.e., the \bar{d}_j are insufficiently extreme. The interpretation is that people tend to be overconfident in their judgments. Much of the earlier research establishing this result has been reviewed by Lichtenstein, Fischhoff, and Phillips (1982) and by Wallsten and Budescu (1983). More recent research has been reviewed by Erev, Wallsten, and Budescu (1994), Keren (1991), McClelland and Bolger (1994), and to a lesser extent by Wallsten (1996).

Overconfidence is not a universal finding, and some of the conditions that moderate it have been established (see the prior references). Thus, well-practiced experts, such as US National Weather Service forecasters issuing precipitation probabilities, tend to be well calibrated (although they slightly overpredict rain). Overconfidence decreases, even verges to underconfidence, as problem difficulty decreases. Theories of the judgment process (e.g. Smith and Ferrell, 1983; Gigerenzer, Hoffrage, and Kleinbölting, 1991; Griffin and Tversky, 1992; Juslin, Olsson, and Björkman, 1997; Tversky and Koehler, 1994) attempt to account for the established patterns of over- and underconfidence.

We will focus here on our model of individual estimation processes, because it underlies our research on combining estimates. The primary foundation comes from Erev, Wallsten, and Budescu (1994). That work summarized a number of empirical trends in the subjective probability literature, and noted that they can be interpreted jointly as suggesting that individuals' probability estimates reflect their personal underlying judgment perturbed by a random error component. Erev *et al.* showed that this simple and weak assumption in conjunction with two others predict all the trends they summarized. Specifically, Erev *et al.* assumed that (1) regardless of how people form their judgments, they ultimately assign each event an internal categorical level of confidence; (2) they map confidence categories monotonically to overt discrete responses in a fashion that depends on the nature and parameters of the particular task; and (3) random error is involved at either the confidence or the response level, or at both.

Individual differences in knowledge and opinions are easily accommodated in this approach. Note that Erev *et al.* did not assume that all individuals assign each statement or forecast to the same category. They assumed only that category labels are monotonic with levels of confidence, and that regardless of how opinions are formed and expressed, the process contains random components. They expressed their model in one mathematical form for the purpose of demonstrating how it handles the empirical effects, while Budescu *et al.* (1997) demonstrated the same phenomena with alternative forms. The specific models, however, are not crucial to our main points, which rest only on the underlying qualitative points.

Pfeiffer (1994) provided a somewhat different demonstration that apparent overconfidence can be understood as a possible byproduct of a random component in the system, and Björkman (1994) discussed the necessity of including an error mechanism in one's theory. The point seems already well accepted and models of the judgment process are being reformulated to include error components (Juslin *et al.*, 1977; Soll, 1995). Recent research demonstrates that even when analyzing judgment data in a manner that accounts for this problem, probability estimates tend to be too extreme (Brenner *et al.*, 1996; Budescu *et al.*, 1996; Dawes and Mulford, 1996).

To summarize this section, we believe that the weak, qualitative assumptions specified by Erev *et al.* provide the proper framework for considering how to combine judgments. Under these assumptions, individuals may have different opinions about unique events. Each person accords each event a level of confidence, assigns responses monotonically to confidence categories, and the whole process is subject to random error. Given this model of the individual, our task is to explain why mean estimates tend to decrease CI or CI' for exchangeable and increase DI or DI' for unique events. More generally, we seek to understand the conditions under which taking central tendencies such as simple averages is an optimal or near-optimal combination rule, as well as the conditions under which alternative rules are superior.

THEORY

Relying on the approach of Wallsten and Diederich (1996), we begin by expressing the Erev *et al.* assumptions in very general formal terms and then take up the issue of interjudge correlations. Our model assumes that individuals assign events to categorical levels of confidence, but allows the number of categories to vary over judges. Let $N_k + 1$ denote that number⁸ for judge k ($k = 1, \dots, K$) and denote the individual categories by judge k 's targeted proportion of true events in each, $u_{k0}, u_{k1}, \dots, u_{kN_k}$, where without loss of generality we assume $0 < u_{k0} \leq u_{k1} \leq \dots \leq u_{kN_k} < 1$. Thus

⁸ We use $N_k + 1$ instead of N_k to denote the number of categories used by judge k only to keep this presentation identical to Wallsten and Diederich's, where categories are indexed by $l = 0, 1, \dots, N_k$. To do otherwise would be to invite confusion on the part of readers who consult both papers.

individual k assigns event j ($j = 1, \dots, J$), by whatever cognitive process, to a category and we denote the result by $u_k^{(j)}$, with $u_k^{(j)} \in \{u_{k0}, u_{k1}, \dots, u_{kN_k}\}$. Now we can let U_k be the discrete random variable that takes on values u_{kl} with a distribution depending on the particular J events that judge k assesses. Further, let E_k be a random variable with mean 0 and variance σ^2 , assuming values e_{ik} , representing judge k 's error on trial i for $i = 1, \dots, I$. As the notation implies, we assume the error term is independent of the event being evaluated. Finally, with h being a real-valued function, let R_k be a random variable taking on values r_{ijk} in $(0,1)$, representing judge k 's overt estimate of event j , $j = 1, \dots, J$, on trial i , such that

$$r_{ijk} = h(u_k^{(j)}, e_{ik}), \quad (7a)$$

or more generally,

$$R_k = h(U_k, E_k) \quad (7b)$$

with h increasing in both its arguments.

Equation (7) is a very weak model stating that a judge's probability estimate for an event depends on the confidence category to which he or she assigns it as well as on an independent error factor. Note that the model does not assume the judge is 'reading out' a covert confidence level with additive error. Rather, it only assumes increasing categorical confidence levels, which can be identified by the proportion of events targeted to each level that are true, and that error is somehow involved in the process of overtly estimating the event probability given its assigned category.⁹ Moreover, judges may differ in their numbers of covert confidence categories. Equation (7) subsumes all the models used by Erev *et al.* (1994) and Budescu *et al.* (1996) as well as the stochastic judgment model of Wallsten and González-Vallejo (1994).

In addition to modeling the individual, we also must classify contexts, or problem types, according to expected levels of dependencies among the estimates provided by different judges. Exhibit 1 illustrates such a classification within a framework provided by equation (7b). Each of the numbered cells represents a different set of assumptions about U_k and E_k , and therefore implies different relations among the R_k . The first row in the table denotes contexts in which all judges have precisely the same underlying opinion about each event, j , in the domain. The other rows denote all remaining contexts, for which there is some level of independence or dependence among the U_k . It is useful in these cases to consider these levels conditional on the state of the event being judged. To facilitate doing so, we introduce an indicator variable, S_j , with $S_j = 1$ when event j is true or occurs and $S_j = 0$ otherwise. The simplest case of non-identical U_k , shown in the next complete row of the table, assumes that judges' confidence categories of true events are conditionally independently distributed and that the same holds for false events. The two remaining rows show two other levels of conditional dependence, although more levels could be specified. Finally, we assume the E_k are unconditionally independent and identically distributed over trials within judges. We distinguish, however, whether they are (Yes) or are not (No) identically distributed over judges.

Cells 1 and 2 represent cases in which judges always assign events to the same confidence categories. The cells differ according to whether (cell 1) or not (cell 2) the associated error processes are assumed to

⁹ Note one obvious and one subtle distinction between the present representation and that used by Erev *et al.* (1994) and by Budescu *et al.* (1977). The obvious difference is that we must index individual judges, while the other representations have no need to. The subtle one is that we do not need the notion of *true judgment* and instead index categories only by the targeted proportion of events that are true in each one.

Exhibit 1. A framework for considering patterns of interjudge correlations

U_k identically distributed for all K judges	E_k identically distributed for all K judges	
	Yes	No
Yes	1	2
No		
Conditional on $S_j = 1$ or $S_j = 0$, the U_k are:		
Independent	3	4
Pairwise independent	5	6
Dependent	7	8

be identical. These cells apply, and conceivably are restricted to, situations in which individuals estimate probabilities of mutually defined exchangeable events on the basis of common information. Perhaps there are some special contexts involving probability estimation of unique events to which the cells also apply, but such situations are rare, at best. Wallsten and Diederich added to the model of equation (7) the assumption that the R_k have finite means and finite variances and then used a form of the law of large numbers (Etemadi, 1983) to obtain a natural result for cell 2, and therefore by implication also for cell 1. That is, referring back to equation (7) and defining the mean of K judges' estimates of event j as $M_K^{(j)} = [\sum_k h(u^{(j)}, E_k)]/K$, with $u^{(j)} = u_1^{(j)} = u_2^{(j)} = \dots = u_K^{(j)}$, Wallsten and Diederich showed that $M_K^{(j)}$ approaches a limit that depends on $u^{(j)}$ as K increases. What that limit is depends on additional assumptions that might be invoked to specify a particular model. For example, under the simple (and generally wrong) special case of equation (7b), $R_k = U_k + E_k$ with E_k unbiased, $M_K^{(j)} \rightarrow u^{(j)}$ as K increases. In terms of the indices presented above, if $u^{(j)}$ equals the long-run relative frequency for event j , $j = 1, \dots, J$, then CI and CI' approach 0 as K increases. Other special cases may yield other limits for $M_K^{(j)}$ and therefore for CI and CI' .

Cells 3–8 allow judges to place a given event in different confidence categories and therefore are appropriate for situations in which probability estimates depend on individual knowledge bases and strategies. We assume these cells apply to contexts in which people are estimating the probabilities of unique events. Cells 7 and 8 are probably most representative of real-world situations, in that they assume that judges have correlated but generally different underlying opinions. In contrast, cells 3–6 are somewhat idealized, in that they assume full or pairwise independence among the U_k . Understanding the combination problem in these latter cases is a first step to understanding it in the more complex ones.

In fact, Wallsten and Diederich (1996), making somewhat stronger assumptions than they used in considering cells 1 and 2, achieved a very powerful result for cell 6 (and by implication for cells 3–5, as well), which we now describe. We present their assumptions and theorem informally here and reproduce them formally in the appendix below. We also provide the intuition behind the proof here, as it is rather instructive. The assumptions for all k are:

- (1) Equation (7) holds, with the restriction that N_k is even.
- (2) R_k are pairwise independent random variables, conditional on event j being true or false, each having finite mean and finite variance.
- (3) The confidence categories, $l = 0, 1, \dots, N_k$, are symmetric about 0.5 and their probabilities of use are symmetric about the central category.
- (4) The errors, e_{ik} , are such that expected responses conditional on u_{kl} regress to 0.5 in a manner that is symmetric about 0.5.

In addition to taking equation (7) as the basic model, assumption 1 stipulates that judges have an odd number of confidence categories. In conjunction with parts of assumptions 3 and 4 (see the Appendix), the consequence is that judges are assumed to have a central confidence category located at 0.5. Although this requirement may be psychologically reasonable, it was made for mathematical convenience. It is in fact innocuous because assumption 3 allows a 0 probability of using the central category. More generally, assumption 3 is that judges set up and use their confidence categories in symmetric fashions. Assumption 4 is that error is independent of the state of the event being judged, regressive relative to the confidence category and symmetric about 0.5. Finally, assumption 2 places us in cell 6 of Exhibit 1 and invokes mild structural conditions. These assumptions, although stronger than those used for cell 2, in fact do little to restrict the cognitive models of individual judges to which the resulting theorem applies.

Wallsten and Diederich's (1996) two-part theorem is: If assumptions 1–4 hold, then for any event, j :
 (a) The mean expected probability estimate of K judges equals a value greater than 0.5 if the statement is true and a value less than 0.5 if it is false. Expressed differently,

$$P(E_K^{(j)} > 0.5 | S_j = 1) = P(E_K^{(j)} < 0.5 | S_j = 0) = 1$$

where $E_K^{(j)}$ is the mean expected estimate of K judges. Employing this result in Bayes' Rule and also making use of a form of the law of large numbers (Etemadi, 1983) yields the second part of the theorem: (b) As the number of judges, K , increases, the probability that an event is true approaches 1 for mean actual estimates greater than 0.5 and approaches 0 for mean actual estimates less than 0.5. That is,

$$\text{as } K \rightarrow \infty, \quad P(S_j = 1 | M_K^{(j)}) \rightarrow \begin{cases} 1 & \text{if } M_K^{(j)} > 0.5 \\ 0 & \text{if } M_K^{(j)} < 0.5 \end{cases} \quad (9)$$

The empirical implication of this result is that if assumptions 1–4 hold, then of all events with mean estimates greater (less) than 0.5, the proportion that are true (false) should approach 1 as K increases. In terms of the indices presented above, DI increases to a limit and DI' increases without bound as K increases.

Upon first consideration, this is a remarkable result. The ideal is achievable! Merely by averaging together a sufficient number of pairwise independent estimates that are possibly perturbed by error and are monotonically related to outcome relative frequency (which roughly describes the judgments of most people operating in domains for which they have some knowledge), one can determine the truth or falsity of any statement with certainty. Of course, we cannot expect to observe that result generally. But a careful inspection of the proof is very instructive regarding the conditions under which it might be approximated empirically and the ways in which the system should be modified to make it more general and realistic.

The structure of Wallsten and Diederich's proof is as follows. First, they prove that if assumption 3 holds, then the proportions of true and false events in the universe under consideration each equal 0.5. In other words, if it is not the case that half the events are true and half false, then assumption 3 must be violated. This result is not as strong as it seems because any reasonably structured domain will be closed with respect to complementation.¹⁰ Next, assumptions 1–3, and 4 imply that although a given judge may place a given event in any confidence category, on average true events will fall in

¹⁰ For example, any set of statements for which judges are estimating probabilities (e.g. statements of the sort 'A is true') implicitly or explicitly also contains their complements ('A is false').

categories above, and false events will fall in categories below, the central one. This is the key point. As a consequence of it, the mean expected estimate over all judges is greater than 0.5 for true events and less than 0.5 for false ones. Bayes' Rule then provides the expression for the probability that an event is true given its expected mean estimate. Finally, to complete the proof, assumption 2 provides the condition for a law of large numbers to apply (Etemadi, 1983), which guarantees that as K increases the observed mean estimate approaches the expected value. Therefore, the probability an event is true approaches 1 or 0 according to whether the mean is above or below 0.5.

Thus, the law of large numbers applies here, just as it does in cells 1 and 2, where underlying judgments are equal except for an error term (or as it might when observers are estimating physical dimensions). But the convergence is not to a particular underlying confidence value. Rather, it is to a population mean estimate that varies according to the state of the event being judged. Our assumptions guarantee that these mean estimates are greater for true than for false events.

If this result does not apply in general, what is wrong with the assumptions? For one thing, assumptions 3 and 4 cannot in the end be correct. In addition to the result just described, these assumptions also imply that the distribution of estimates over judges is the same for all true and for all false events, respectively, and moreover that these two distributions are mirror images of each other. In other words, assumptions 3 and 4 imply that all events are equally easy or difficult to judge. Thus, for example, with regard to latitudes of cities of the world, the distributions of judges' probability estimates are predicted to be identical for the true statements: 'New York is south of Rome', 'Paris is north of Rio de Janeiro', and 'Tokyo is north of Los Angeles'. Of course, that result will not obtain with most groups of judges. On the other hand, assumptions 3 and 4 may be correct, or nearly so, for many populations of judges with respect to many domains of interest.

A second problem is that assumption 2, conditional pairwise independence, will not generally hold. As Yates pointed out in reviewing an earlier draft of this article, '... it is hard to imagine how averaging can give rise to any improvement in ... "diagnosticity" unless it somehow leads to a better exploitation of the actual diagnosticity that exists in the environment'.¹¹ It is precisely the assumption of conditional pairwise independence that allows this exploitation to occur. Judges differ in their knowledge as well as in the cues and strategies that they use, and therefore in the features they treat as diagnostic, when forming probability estimates. Averaging improves the diagnostic value of the results only to the degree that judges behave in a conditionally uncorrelated fashion. In fact we do not generally know to what extent judges' probability estimates are conditionally related, nor do we know the population or domain characteristics that affect such relationships. There may be knowledge domains and populations of judges for which assumption 2 is reasonable. Finally, at a theoretical level we do not know how robust the result is to violations of this assumption.

To summarize this section, when K judges' probability estimates are based on identical underlying opinions, such as when they use common information to estimate the probabilities of well-defined exchangeable events, very weak assumptions imply that the mean of the estimates of event j converges to a value unique for that event as K increases. Otherwise, we distinguish whether events are true or false, as well as levels of independence among judges' estimates conditional on the event state. Relatively weak conditions on the variables in equation (7) in conjunction with an assumption of conditional pairwise independence among the probability estimates leads to the remarkable result that the mean of K estimates of a true (or false) event converges to a value greater (or less) than 0.5 as K increases. Consequently, the probability that the event is true (or false) approaches 1 (or 0) as K increases. While this result will not hold universally, its empirical and theoretical limits remain to be explored. We now turn to empirical tests of the developments outlined here.

¹¹ Personal correspondence from J. F. Yates dated 26 April 1996.

DATA

We reanalyzed data from two studies that bear, respectively, on cells 2 and 6 of Exhibit 1. We take up the former first.

Cell 2

One of Erev and Wallsten's (1993) experimental conditions appears to meet the cell 2 (maybe even the stronger cell 1) requirement that the judgments underlying the individual subjects' probability estimates be identical except for an error component. Sixty subjects observed computer displays from which they could infer relative frequency information about 36 distinct events (which among them had 26 different objective probabilities), and provided subjective probability estimates of those events. Erev *et al.* (1994) had shown that one could use these data to conclude that respondents were either over- or underconfident, depending on the method of analysis. Individuals appeared overconfident when, as is commonly done in calibration research, the expected relative frequencies of occurrence, \bar{d}_j , were analyzed as a function of subjective probability categories, f_j . They appeared underconfident when the contingencies were reversed, as is commonly done in research comparing subjective estimates to objective values based on expected relative frequency or Bayesian calculations.

Erev (1990/1991) reported the mean correlation between individual sets of estimates and the objective probabilities as 0.72 for these data. We now have analyzed the data further and, not surprisingly, found the pairwise correlations between individuals' subjective estimates to be high (median correlation = 0.60 with the central 50% of the correlations ranging from 0.45 to 0.72). Assuming this situation falls in Cell 2, then, as indicated earlier, on the basis of law of large number considerations, the means of the 60 judgments for event, $a_j, j = 1, \dots, 36$, should converge to a limit, L_j . If, in addition, we assume (1) that each event is assigned an underlying confidence category equal to its objective probability, i.e. $u_k^{(j)} = P(a_j)$ for all k and j , and (2) an error function regressive towards 0.5,¹² then L_j should be close to $P(a_j)$, but somewhat anti-regressive. That is, when plotted as a calibration curve, the points should lie close to the diagonal but slightly below it for $f_j < 0.5$ and slightly above it for $f_j > 0.5$. Moreover, both CI and CI' should decrease with averaging.

Exhibit 2 illustrates the results of averaging the 60 estimates per event. The abscissa shows subjects' probability estimates divided into 11 categories centered at 0.025, 0.10, 0.20, \dots , 0.90, 0.975. The actual values plotted are the means of the estimates in each category, f_j . The ordinate shows the relative frequencies of the events conditional on the estimates, \bar{d}_j . The $K = 1$ function is the average of the individual calibration curves (it is identical to the S-curve of Erev *et al.*'s Fig. 2) and the $K = 60$ function is the calibration curve of the averaged judgments. As expected, the relative frequencies conditional on the mean judgments are much closer to the diagonal than are those conditional on the individual estimates. Moreover, the points have moved from extreme overconfidence to slight underconfidence with averaging.

Exhibit 3 shows various comparative indices for $K = 1$ and $K = 60$. The first set are calculated on the data conditioned on the estimates and correspond to the graphs in Exhibit 2. First is the accuracy criterion, CI , which we expect to and which does improve substantially with averaging. Next, we show DI , which also evidences some improvement. Thus in this case, averaging improves the probability score, \overline{PS} , from 0.232 to 0.212 (see equation (2)) by reducing both CI and DI .

The following two columns of Exhibit 3 show the square roots of CI and DI , which are more easily interpreted than the indices themselves, because they revert back to the original metric of the estimates. The next column provides the measure of over- or underconfidence that Erev *et al.* (1994) defined for

¹² Any reasonable error function for estimates in the closed [0,1] interval will have this property.

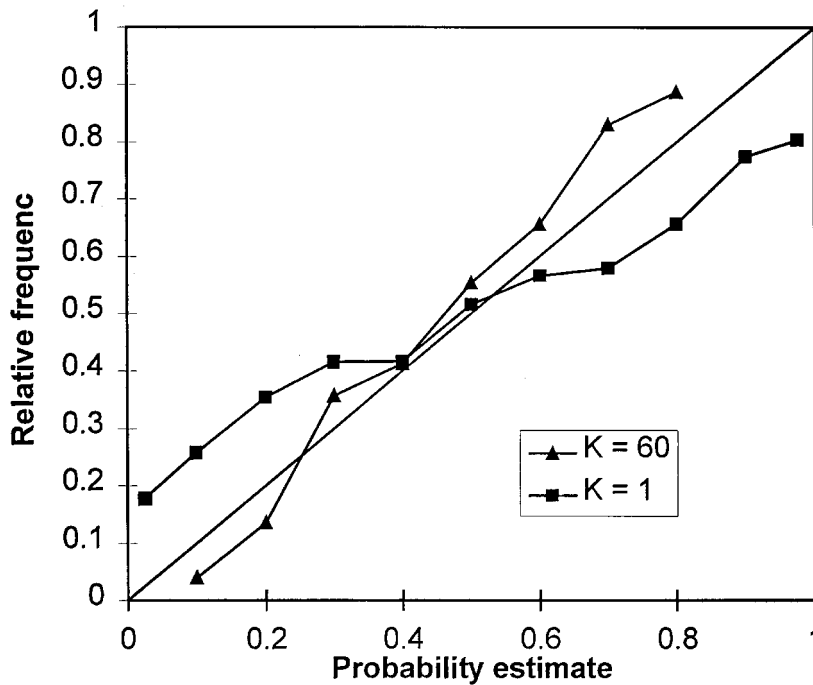


Exhibit 2. Event relative frequency as a function of subjective probability category for the data from Erev and Wallsten (1993) for individual subjects ($K = 1$) and for probability estimates averaged for each event over the whole group ($K = 60$)

Exhibit 3. Accuracy and overconfidence indices for the Erev and Wallsten (1994) data

K	Conditioned on the estimates					Conditioned on the objective probabilities	
	CI	DI	$CI^{0.5}$	$DI^{0.5}$	$CONF_S$	CI'	$(CI')^{0.5}$
1	0.016	0.033	0.126	0.181	0.139	0.059	0.244
60	0.006	0.043	0.075	0.208	-0.042	0.011	0.102

data analyzed conditional on the estimates, $CONF_S$. It, too, is in the metric of the estimates, and is defined as the weighted mean directional deviation of the calibration points from the diagonal, constructed so that positive values indicate over- and negative values underconfidence. Specifically, in the notation of this paper,

$$CONF_S = \frac{1}{N} \left[\sum_{f_j < 0.5} N_j(\bar{d}_j - f_j) + \sum_{f_j > 0.5} N_j(f_j - \bar{d}_j) \right]$$

$CONF_S$ provides a measure of the change from considerable overconfidence to mild underconfidence.¹³

¹³ Erev *et al.* defined a corresponding index, $CONF_O$, for analyses conditioned on the objective values. It consistently shows underconfidence for these data.

Finally, because it is more natural to analyze estimates of exchangeable events by conditioning on the events than on the estimates, Exhibit 3 also gives CI' and its root. CI' based on $K = 60$ is considerably smaller than the mean CI' based on $K = 1$. DI' is not defined for this paradigm.

Cell 6

In Wallsten, Budescu, and Zwick's (1993) study, 21 subjects (primarily undergraduates) gave verbal and numerical probability estimates that general knowledge statements were true. We restrict attention here to the numerical estimates. Assuming that subjects brought different backgrounds to the statements they were judging, this study may approximate the condition of cell 5 or 6, which requires that the judgments underlying the probability estimates are conditionally pairwise independent. If the condition is met exactly, the conditional pairwise correlations among the subjective probability estimates will be 0.

In the numerical condition, subjects judged 75 true and 75 false statements in one session and on a separate occasion they judged the, respectively, false and true complements. The sums of the probability estimates of the complementary statements were very close to and did not deviate significantly from 1. Therefore, the estimates of the true statements plus one minus the estimates of the false ones can be considered replicates. The within-subject reliability correlations on these replicated estimates are similar in value to the pairwise between-subject correlations of the Erev and Wallsten (1993) data (median of 0.63 with the central 50% range from 0.53 to 0.71). In contrast, the median of the conditional pairwise between-subject correlations is 0.26 and the central 50% ranges from 0.16 to 0.39. Thus, the assumption required for cell 6 is approximated, but clearly not met, and we have an opportunity to test the robustness of the predicted result to violations of conditional pairwise independence.

Consistent with the calibration literature, Wallsten *et al.* (1993) found that relative frequencies of true statements given the probability judgments were insufficiently extreme, leading to the conclusion that the subjects were overconfident. Under the theorem outlined above, we expect that in the limit the mean estimates for all false statements converge to a value less than 0.5 and those for all true statements to a value greater than 0.5. Consequently, the probability that a statement is true approaches 0 given a mean estimate less than 0.5 and approaches 1 given a mean estimate greater than 0.5 as the number of judges increases.

Exhibit 4 shows the mean calibration curve for individual subjects ($K = 1$ — combining the numerical curves from Wallsten *et al.*'s Figures 3 and 4) as well as the calibration curve based on averaging the 21 estimates per statement ($K = 21$). Consistent with expectations, the curve becomes more extreme and approximates a step function as K increases. Over all items, the probability that a statement is true given mean estimates less than 0.5 is 0.12 and given mean estimates greater than 0.5 is 0.84.

Exhibit 5 summarizes the various indices. As expected, CI does not improve with averaging. Rather, it increases as the data points depart from the diagonal. DI , however, does improve to a substantial degree. Consequently, the overall probability score, \overline{PS} , improves from 0.213 to 0.153 with averaging. Conditioning on the state of the events rather than on the estimates, DI' almost doubles. CI' is not defined in this paradigm.

Given the considerable increase in diagnosticity of the estimates with averaging even in the presence of less than perfect pairwise independence, it is of interest to ask how the indices improve with K . For simplicity, we restrict attention to DI' . To answer the question, we randomly eliminated one subject in order to have a number of subjects (20) that could easily be broken into subsets of different sizes. Then we took all subsets of size 5 and 10, averaged the probability estimates per statement within the subsets, calculated DI' , and took the mean value within each subset size. In addition, we calculated mean DI' at $K = 1$ and DI' at $K = 20$ for those 20 subjects. The results are 1.01, 1.54, 1.68, and 1.76, respectively

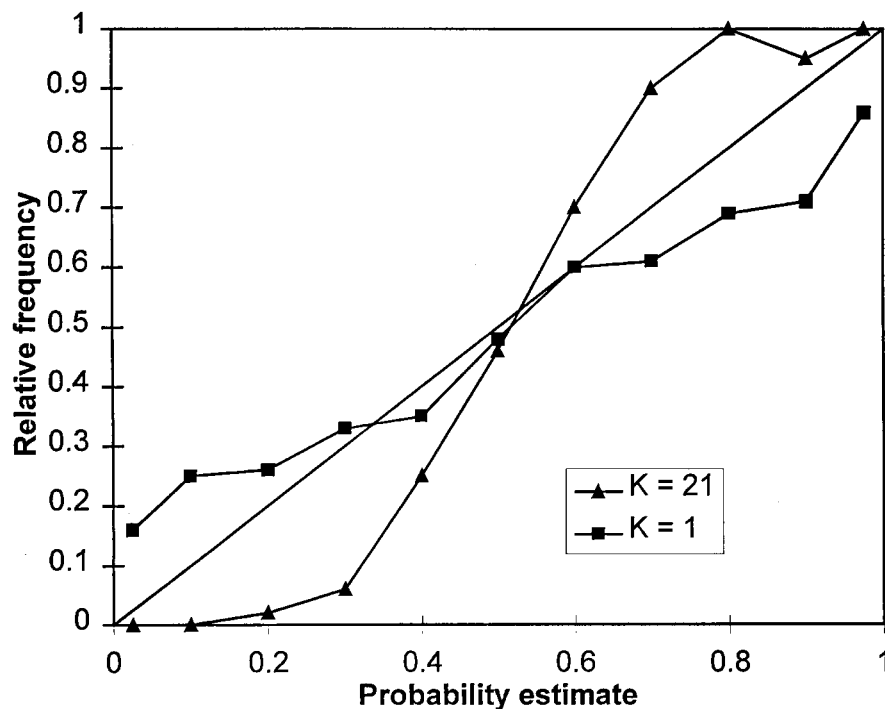


Exhibit 4. Relative frequency correct as a function of subjective probability category for the data from Wallsten, Budescu, and Zwick (1993) for individual subjects ($K = 1$) and for probability estimates averaged for each statement over the whole group ($K = 21$)

for, $K = 1, 5, 10,$ and 20 . Thus, substantial improvement was obtained by averaging over as few as five subjects.

DISCUSSION

We first will discuss the theoretical and the empirical results in turn, and then some of the remaining issues. Finally, we will draw some broad conclusions.

Theoretical results

Our strategy for considering how to combine probability estimates relies on (1) developing a weak and therefore very broadly applicable model of the individual judge and (2) classifying probability estimation contexts. We believe our results prove that the approach is successful. Our model of the

Exhibit 5. Accuracy and overconfidence indices for the Wallsten *et al.* (1993) data

K	Conditioned on the estimates					Conditioned on the objective probabilities
	CI	DI	$CI^{0.5}$	$DI^{0.5}$	$CONF_S$	DI'
1	0.010	0.047	0.098	0.181	0.088	0.99
21	0.024	0.117	0.154	0.342	-0.161	1.81

judge is embodied in equation (7), in conjunction with the associated constraints on the variables. This equation provides the basis for classifying contexts in Exhibit 1. For all practical purposes, cells 1 and 2 of Exhibit 1 apply to subjective probability estimates of well-defined exchangeable events, which we assume judges form on the basis of common information, generally perceived relative frequencies. Cells 3–8 apply to estimates of unique events, which we assume judges base on their individual stores of knowledge and personal understanding. It is useful in these latter cells to consider the estimates conditional on event state (true or false) and to distinguish various levels of conditional dependence. Within this framework, we have achieved a good understanding of the consequences of averaging subjective probability estimates in cells 1–6 and are in a position to draw some useful generalizations.

First, it is useful to treat the exchangeable and unique cases separately when possible, because the effects and consequences of averaging probability estimates are not the same in the two cases. Although some may consider the point obvious, to our knowledge it has not been made explicitly before.

When dealing with exchangeable events (cells 1 and 2), it is useful to consider the estimates conditional on the event probabilities or long-run relative frequencies. According to Wallsten and Diederich (1996), as K increases, the mean estimate of an event should approach a limit that depends on the event probability. Assuming that limit is not too far from the probability, then CI' will decrease as the number of judges, K , increases. Similarly, if the data are treated conditional on the estimates (perhaps because there is not agreement on the definition of the exchangeable events), CI also will decrease with K . Our reanalysis of the Erev and Wallsten data confirms these expectations.

Quite a different result is predicted when K judges estimate the probabilities of unique events (cells 3–6), there is reason to believe their underlying confidence categories are pairwise independent, and certain symmetry conditions hold. In these cases, according to Wallsten and Diederich the probability accorded an event ought to approach 0 or 1 as K increases, depending on whether its mean estimate is below or above 0.5. Consequently, DI' and DI will both improve with K , but (in the limit) CI will worsen. Our reanalysis of the Wallsten *et al.* data yielded precisely these results. Note that some authors (e.g. Winkler and Poses, 1993) have found CI to improve with averaging.

At this point we do not know the implications of averaging multiple estimates for the more realistic conditions covered by cells 7 and 8. At the extreme, cells 7 and 8 wrap back to cells 1 and 2 (identical underlying confidence categories per event over judges), but we would guess that such an extreme condition rarely if ever exists. From the opposite perspective, our empirical results, which we will discuss in more detail below, suggest that the diagnostic impact of averaging estimates is robust with respect to departures from pairwise independence.

The framework proposed here can help resolve controversies in the probability judgment literature. Consider the unanimity principle (Morris, 1983) mentioned earlier, which has been much discussed by decision theorists (i.e. Lindley, 1986; Morris, 1986; Winkler, 1986). Lindley provided an example in which a decision maker properly violates this principle when she uses Bayes' Rule to update her own probability about an event given multiple identical estimates by others. Morris responded that if the unanimous probability estimates are of a Bernoulli process, then according to a Bayesian analysis, the unanimity principle holds. And Winkler said that one cannot apply a single principle to all cases, but must model each situation appropriately. Wallsten and Diederich's (1996) results suggest that all three theorists are correct. The principle does hold for the case of exchangeable events of the sort used by Morris in his illustration, but not for those of unique events covered by cells 3–6, which is where Lindley's example falls. That is, if for some reason, multiple judges give the same estimate for a particular exchangeable event, then that value is likely a very good estimate of the event's true probability. In contrast, if multiple judges give the same estimate for a particular unique event and there is reason to believe their underlying confidence categories are approximately conditionally pairwise independent, then the probability accorded that event ought to approach 0 or 1 depending on whether the mean estimate is below or above 0.5.

Note, finally, that the Wallsten and Diederich theorem does provide a Bayesian pooling procedure, but one that is different from the other Bayesian methods mentioned above. In the spirit of Lindley's (1982) frequency calibration model, the theorem provides (when the assumptions are met) a probability that an event is true given a pool of estimates, but only in the limit. For any finite sample of estimates, this probability approaches 0 or 1, depending on whether the mean of the probability judgments, $M_K^{(j)}$, is less than or greater than 0.5 (see equation (9)). At this time, we have no way to estimate that conditional probability except empirically.

Empirical results

The data strongly substantiate Wallsten and Diederich's (1996) analysis and therefore the cognitive model on which it is based. Perhaps of greatest importance the empirical results confirm that the consequences of averaging estimates are different for exchangeable and unique events. In the one case, the estimates become more accurate and in the other they become more diagnostic.

In the exchangeable case, the mean estimates are closer to the event probabilities for $K = 60$ than for $K = 1$. Moreover, as expected, the mean estimates are slightly regressive relative to those probabilities. Although we did not use various sized subsets of the data to check convergence, it is apparent that on average the accuracy indices will improve as K increases.

Perhaps the more surprising result is the one that occurred with the estimates of unique events. Subjects' estimates of true and false statements, respectively, were weakly correlated. Nevertheless, the means of the 21 estimates per statement were considerably more diagnostic than were the individual estimates. On theoretical grounds, we predicted averaging to improve the diagnosticity of the estimates when they are conditionally pairwise independent, but we did not know what to expect if this assumption does not strictly hold. It is clear, and of potential practical import, that the outcome appears to be robust with respect to violations. Also of practical interest is the fact that we gained substantial improvement in diagnosticity by averaging as few as 5 sets of estimates.

Remaining issues

Taken together, the present results provide considerable support for the approach and the underlying theory. But much remains to be done. With regard to estimates of exchangeable events, in view of the fact that the mean estimates appear to converge on slightly regressive values as K increases, it is of interest to find particular transformations, h in equation (7), that describe how individuals convert underlying judgments to overt estimates. Given the correct transformation, one can employ its inverse, h^{-1} , prior to taking averages over judges. In that sense, arithmetic means of subjective estimates may not be the most efficient.

Interestingly, the particular response transformation may be less important when considering probability estimates of unique events under conditions that approach conditional pairwise independence. This is because averaging increases diagnosticity rather than accuracy in such cases. The point remains to be explored theoretically and with suitable computer simulations. In a similar fashion, simulations will be helpful in assessing exactly how robust Wallsten and Diederich's (1996) theorem is to violations of conditional pairwise independence, as well as in assessing the rate of convergence as K increases.

The major open problem concerns cells 7 and 8, for which we currently have no results. It may turn out the violations of Wallsten and Diederich's assumptions 3 and 4 are more problematic than are violations of their assumption 2. We expect to move into cells 7 and 8 while simultaneously overcoming this problem by adopting a restrictive special case of equation (7). One candidate is the Stochastic Judgment Model (Wallsten and González-Vallejo, 1994), which makes specific assumptions about distributions of true and false events (or statements) and about the individual judgment process.

SUMMARY AND CONCLUSIONS

Our goal is to develop and evaluate principled procedures for combining subjective probability judgments of categorical events for the purpose of increasing their diagnostic value. We claim that the principles from which the combination rule is derived must come from suitable models of the judges' cognitive processes, as well as the structure of the information base. The research presented here demonstrates that very weak assumptions about the judge, which are consistent with considerable data at the individual level, are sufficient to yield strong, even counter-intuitive predictions, and to explain the available results on the averaging of probability estimates. The model, as we have developed it thus far, seems valid in some contexts but may not apply to others, in that it fails to allow differences among properties of events and judges. We believe that we can overcome these limitations by building on a more structured model of individual judgment processes.

APPENDIX

To aid the reader, this appendix reproduces the formal statement of Wallsten and Diederich's (1996) assumptions and theorem for cell 6 of Exhibit 1. Consult the original source for the formal proof.

Assume for $k = 1, \dots, K$:

- (1) Equation (7) holds, with the restriction that N_k is even.
- (2) R_k are pairwise independent random variables, conditional on $S_j = 1$ or $S_j = 0$, each having finite mean and finite variance.
- (3) The values u_{kl} of U_k have the following properties:
 - (a) They are symmetric about 0.5, i.e. $u_{kl} + u_{k, N_k - l} = 1$ for $l = 0, \dots, N_k/2 - 1$, $u_{k, N_k/2} = 0.5$;
 - (b) The probabilities are equal for symmetric pairs, i.e. $P(U_k = u_{kl}) = P(U_k = u_{k, N_k - l})$ for $l = 0, \dots, N_k/2 - 1$; with, of course

$$0 \leq P(U_k = u_{kl}) < 1 \quad \text{and} \quad \sum_{l=0}^{N_k} P(U_k = u_{kl}) = 1.$$

- (4) The expected responses conditional on a given confidence category over trials, $E(R_k | U_k = u_{kl}, S_j = 1)$ and $E(R_k | U_k = u_{kl}, S_j = 0)$, have the following properties:
 - (a) They are regressive with respect to u_{kl} , i.e.

$$E(R_k | U_k = u_{kl}, S_j = 1) = E(R_k | U_k = u_{kl}, S_j = 0) = 0.5w_{kl} + (1 - w_{kl})u_{kl}, \quad 0 < w_{kl} < 1;$$

- (b) They are symmetric about 0.5, i.e.

$$\begin{aligned} E(R_k | U_k = u_{kl}, S_j = 1) + E(R_k | U_k = u_{k, N_k - l}, S_j = 1) \\ = E(R_k | U_k = u_{kl}, S_j = 0) + E(R_k | U_k = u_{k, N_k - l}, S_j = 0) = 1 \\ \text{for } l = 0, \dots, \frac{N_k}{2} - 1; \quad \text{and} \\ E(R_k | U_k = u_{k, N_k/2}, S_j = 1) = E(R_k | U_k = u_{k, N_k/2}, S_j = 0) = 0.5. \end{aligned}$$

We require one additional piece of notation to present the theorem. Recall the definition of the mean of K judges' probability estimates of event j , $M_K^{(j)}$. Analogously, define the mean expected estimate as, $E_K^{(j)} = [\sum_k E(R_k^{(j)})]/K$, where $E(R_k^{(j)})$ is the expected estimate of judge k to event j .

Theorem: If assumptions 1–4 hold, then for a given event j for K judges

- (a) $E_K^{(j)} > 0.5$ if $S_j = 1$ and $E_K^{(j)} < 0.5$ if $S_j = 0$
- (b) as $K \rightarrow \infty$, $P(S_j = 1 | M_K^{(j)}) \rightarrow \begin{cases} 1 & \text{if } M_K^{(j)} > 0.5 \\ 0 & \text{if } M_K^{(j)} < 0.5 \end{cases}$

ACKNOWLEDGEMENTS

This research was supported by National Science Foundation Grants No. SBR-9222159 and SBR-9601281 to Thomas S. Wallsten, NSF Grant No. SBR-9632448 to David V. Budescu, and Deutsche Forschungsgemeinschaft Grant No. Di 506/2-1 to Adele Diederich. We thank Bob Clemen, Claudia González-Vellajo, In Jae Myung, Jack Soll, Frank Yates, and an anonymous reviewer for very thoughtful and helpful comments on earlier versions of this paper.

REFERENCES

- Björkman, M. 'Internal cue theory: Calibration and resolution of confidence in general knowledge', *Organizational Behavior and Human Decision Processes*, **58** (1994), 386–405.
- Brenner, L. A., Koehler, D. J., Liberman, V., and Tversky, A. 'Overconfidence in probability and frequency judgments', *Organizational Behavior and Human Decision Making*, **65** (1996), 212–19.
- Budescu, D. V., Erev, I., and Wallsten, T. S. 'On the importance of random error in the study of probability judgment. Part I: New theoretical developments', *Journal of Behavioral Decision Making*, **10** (1997) 157–171.
- Budescu, D. V. and Wallsten, T. S. 'Subjective estimation of precise and vague uncertainties', in Wright, G. and Ayton, P. (eds), *Judgmental Forecasting* (pp. 63–81), Chichester: Wiley, 1987.
- Bunn, D. W. and Mustafaoglu, M. M. 'Forecasting political risk', *Management Science*, **24** (1978) 1557–67.
- Clemen, R. T. 'Calibration and the aggregation of probabilities', *Management Science*, **32** (1986), 312–14.
- Clemen, R. T. 'Combining forecasts: A review and annotated bibliography', *International Journal of Forecasting*, **5**(4) (1989), 559–609.
- Clemen, R. T. and Murphy, A. H. 'The expected value of frequency calibration', *Organizational Behavior and Human Decision Processes*, **46** (1990), 102–17.
- Clemen, R. T. and Winkler, R. L. 'Calibrating and combining precipitation probability forecasts', in Viertl, R. (ed.), *Probability and Bayesian Statistics* (pp. 97–110), New York: Plenum, 1987.
- Clemen, R. T. and Winkler, R. L. 'Aggregating point estimates: A flexible modeling approach', *Management Science*, **39** (1993), 501–15.
- Cooke, R. M. *Experts in Uncertainty: Opinion and subjective probability in science*, New York: Oxford University Press, 1991.
- Curtis, P. B., Ferrell, W. R., and Hillman, B. J. 'Improved imaging diagnosis by sequentially combined confidence judgments', *Investigative Radiology*, **23** (1988), 342–7.
- Davidson, D., Suppes, P., and Siegel, S. *Decision Making: An experimental approach*, Stanford, CA: Stanford University Press, 1957.
- Dawes, R. M. and Mulford, M. 'The false consensus effect and overconfidence: Flaws in judgment, or flaws in how we study judgment?' *Organizational Behavior and Human Decision Making*, **65** (1996) 201–11.
- De Finetti, B. 'La prévision: Ses lois logiques, ses sources subjectives', *Annales del' Institut Henri Poincaré*, **7** (1937), 1–68, English translation in Kyburg, H. E. Jr and Smokler, H. E. (eds), *Studies in Subjective Probability*, New York: Wiley, 1964.
- De Finetti, B. *Theory of Probability: a critical introductory treatment*, New York: Wiley, 1974.
- Erev, I. *The sensitivity of human behavior to the likelihood of future events, and the information reduction assumption*, Doctoral dissertation, University of North Carolina at Chapel Hill, 1990. *Dissertation Abstracts International*, **51** (1991), 5017.

- Erev, I. and Wallsten, T. S. 'The effect of explicit probabilities on decision weights and on the reflection effect', *Journal of Behavioral Decision Making*, **6** (1993), 221–41.
- Erev, I., Wallsten, T. S., and Budescu, D. V. 'Simultaneous over- and underconfidence: The role of error in judgment processes', *Psychological Review*, **101** (1994), 519–27.
- Etemadi, N. 'On the laws of large numbers for nonnegative random variables', *Journal of Multivariate Analyses*, **13** (1983), 187–93.
- French, S. 'Calibration and the expert problem', *Management Science*, **32** (1986), 315–21.
- Genest, C. and Zidek, J. V. 'Combining probability distributions: A critique and an annotated bibliography', *Statistical Science*, **1** (1986), 114–48.
- Gigerenzer, G., Hoffrage, U., and Kleinbölting, H. 'Probabilistic mental models: A Brunswikian theory of confidence', *Psychological Review*, **98** (1991), 506–28.
- Goldberg, L. R. 'Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences', *Psychological Bulletin*, **73** (1970), 422–32.
- Gordon, K. 'Group judgments in the field of lifted weights', *Journal of Experimental Psychology*, **7** (1924), 398–400.
- Graham, J. R. 'Is a group of economists better than one? Than none?' *Journal of Business*, **69** (1996), 193–232.
- Griffin, D. and Tversky, A. 'The weighing of evidence and the determinants of confidence', *Cognitive Psychology*, **24** (1992), 411–35.
- Heath, C. and Tversky, A. 'Preference and belief: Ambiguity and competence in choice under uncertainty', *Journal of Risk and Uncertainty*, **4** (1991), 5–28.
- Hogarth, R. M. 'Methods for aggregating opinions', in Jungermann, H. and d. Zeeuw, G. (eds), *Decision Making and Change in Human Affairs* (pp. 231–55), Dordrecht: D. Reidel, 1977.
- Hogarth, R. M. 'A note on aggregating opinions', *Organizational Behavior and Human Performance*, **21** (1978), 40–46.
- Howell, W. C. 'Uncertainty from internal and external sources: A clear case of overconfidence', *Journal of Experimental Psychology*, **89**(2) (1971), 240–43.
- Juslin, P., Olsson, H., and Björkman, M. 'Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgments', *Journal of Behavioral Decision Making*, **10** (1997), 189–209.
- Kahneman, D. and Tversky, A. 'Variants of uncertainty', *Cognition*, **11** (1981), 143–57.
- Keren, G. 'Calibration and probability judgments: Conceptual and methodological issues', *Acta Psychologica*, **77** (1991), 217–73.
- Levy, W. B. and Deliç, H. 'Maximum entropy aggregation of individual opinions', *IEEE Transactions on Systems, Man & Cybernetics*, **24** (1994), 606–13.
- Lichtenstein, S., Fischhoff, B., and Phillips, L. D. 'Calibration of probabilities: The state of the art to 1980', in Kahneman, D., Slovic, P. and Tversky, A. (eds), *Judgment under Certainty: Heuristics and biases* (pp. 306–34), Cambridge: Cambridge University Press, 1982.
- Lindley, D. 'The improvement of probability judgments', *Journal of the Royal Statistical Society. Series A*, **145** (1982), 117–26.
- Lindley, D. V. 'Another look at an axiomatic approach to expert resolution', *Management Science*, **32** (1986), 303–6.
- Locke, A. 'Integrating group judgments in subjective forecasts', in Wright, G. and Ayton, P. (eds), *Judgmental Forecasting* (pp. 109–28), Chichester: Wiley, 1987.
- Lorge, I., Fox, D., Davitz, J., and Brenner, M. 'A survey of studies contrasting the quality of group performance and individual performance', *Psychological Bulletin*, **55** (1958), 337–73.
- Marley, A. A. J. 'Aggregation theorems and multidimensional stochastic choice models', *Theory and Decision*, **30** (1991), 245–72.
- McClelland, A. G. R. and Bolger, F. 'The calibration of subjective probabilities: Theories and models 1980–94', in Wright, G. and Ayton, P. (eds), *Subjective Probability* (pp. 453–82), Chichester: Wiley, 1994.
- McNees, S. K. 'The uses and abuses of "consensus" forecasts', *Journal of Forecasting*, **11** (1992), 703–10.
- Meehl, P. E. 'A comparison of clinicians with five statistical methods of identifying psychotic MMPI profiles', *Journal of Counseling Psychology*, **6** (1959), 102–9.
- Morris, P. A. 'Combining expert judgments: A Bayesian approach', *Management Science*, **23** (1977), 679–93.
- Morris, P. A. 'An axiomatic approach to expert resolution', *Management Science*, **29** (1983), 24–32.
- Morris, P. A. 'Observations on expert aggregation', *Management Science*, **32** (1986), 321–8.
- Murphy, A. H. 'A new vector partition of the probability score', *Journal of Applied Meteorology*, **12** (1973), 595–600.
- Murphy, A. H. and Winkler, R. L. 'Diagnostic verification of probability forecasts', *International Journal of Forecasting*, **7** (1992), 435–55.
- Myung, I. J., Ramamoorti, S., and Bailey, A. D. 'Maximum entropy aggregation of expert outcome predictions', *Management Science*, in press.

- National Research Council *Combining Information — Statistical Issues and Opportunities for Research*, Washington, DC: National Academy Press, 1992.
- Pfeiffer, P. E. 'Are we overconfident in the belief that probability forecasters are overconfident?' *Organizational Behavior and Human Decision Processes*, **58** (1994), 203–13.
- Schervish, M. J. 'Comments on some axioms for combining expert judgments', *Management Science*, **32** (1986), 306–12.
- Shafer, G. 'Savage revisited', *Statistical Science*, **1** (1986), 463–501.
- Smith, M. and Ferrell, W. R. 'The effect of base rate on calibration of subjective probability for true-false questions: Model and experiment', in Humphreys, P., Svenson, O. and Vari, A. (eds), *Analyzing and Aiding Decisions* (pp. 469–88), Amsterdam: North-Holland, 1983.
- Soll, J. B. 'Averaging probability judgments as a remedy for overconfidence', paper presented at *Subjective Probability, Utility and Decision Making* (SPUDM-15), Jerusalem, Israel, 1995.
- Tversky, A. and Koehler, D. J. 'Support theory: A nonextensional representation of subjective probability', *Psychological Review*, **101** (1994), 547–67.
- Vesely, W. E. and Rasmuson, D. M. 'Uncertainties in nuclear probabilistic risk analysis', *Risk Analysis*, **4** (1984), 313–22.
- Wallsten, T. S. 'The costs and benefits of vague information', in Hogarth, R. (ed.), *Insights in Decision Making. A tribute to the late Hillel Einhorn* (pp. 28–43), Chicago: The University of Chicago Press, 1990.
- Wallsten, T. S. 'An analysis of judgment research analyses', *Organizational Behavior and Human Decision Making*, **65** (1996), 220–26.
- Wallsten, T. S. and Budescu, D. V. 'Encoding subjective probabilities: A psychological and psychometric review', *Management Science*, **29** (1983), 151–73.
- Wallsten, T. S., Budescu, D. V., and Zwick, R. 'Comparing the calibration and coherence of numerical and verbal probability judgments', *Management Science*, **39** (1993), 176–90.
- Wallsten, T. S. and Diederich, A. 'Proof of the theorem on combining estimates', Working paper, August, 1996.
- Wallsten, T. S. and González-Vallejo, C. 'Statement verification: A stochastic model of judgment and response', *Psychological Review*, **101** (1994), 490–504.
- Whitfield, R. G. and Wallsten, T. S. 'A risk assessment for selected lead-induced health effects: An example of a general methodology', *Risk Analysis*, **9** (1989), 197–207.
- Winkler, R. L. 'Probabilistic prediction: Some experimental results', *Journal of the American Statistical Association*, **66** (1971), 675–85.
- Winkler, R. L. 'Expert resolution', *Management Science*, **32** (1986), 298–303.
- Winkler, R. L. and Murphy, A. H. "'Good" probability assessors', *Journal of Applied Meteorology*, **7** (1968), 751–8.
- Winkler, R. L. and Poses, R. M. 'Evaluating and combining physicians' probabilities of survival in an intensive care unit', *Management Science*, **39** (1993), 1526–43.
- Yaniv, I., Yates, J. F., and Smith, J. E. K. 'Measures of discrimination skill in probabilistic judgment', *Psychological Bulletin*, **110** (1991), 611–7.
- Yates, J. F. 'External correspondence: Decompositions of the mean probability score', *Organizational Behavior and Human Performance*, **30** (1982), 132–56.
- Yates, J. F. 'Subjective probability accuracy analysis', in Wright, G. and Ayton, P. (eds), *Subjective Probability*, Chichester: Wiley, 1994.
- Yates, J. F. and Curley, S. P. 'Conditional distribution analyses of probabilistic forecasts', *Journal of Forecasting*, **4** (1985), 61–73.
- Zajonc, R. B. 'A note on group judgments and group size', *Human Relations*, **15** (1962), 177–80.