ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 30, 132–156 (1982)

External Correspondence: Decompositions of the Mean Probability Score

J. FRANK YATES

University of Michigan

Two evaluative criteria for probabilistic forecasting performance, consistency with the axioms of probability theory and external correspondence with the events that ultimately occur, are distinguished. The mean probability, or Brier score ($\overline{PS}$), is the scoring rule most commonly used to quantify external correspondence. A review is made of methods for decomposing $\overline{PS}$ into components that represent distinct and important aspects of external correspondence. Data from an empirical study of forecasting performance are used to illustrate the interpretation of the components of the most recent decomposition of $\overline{PS}$ (J. F. Yates, *Forecasting performance: A covariance decomposition of the mean probability score*. Paper presented at 22nd Annual Meeting of the Psychonomic Society, Philadelphia, November 1981; also an unpublished manuscript). Substantively, the most important finding of the study was a "collapsing" tendency in forecasting behavior, whereby subjects were inclined to report forecasts of .5 when they felt they knew little about the event in question. This finding is problematic because self-reported knowledge was only minimally related to the actual external correspondence of the subjects' forecasts. A survey of uses of $\overline{PS}$ decompositions suggests, among other things, that current research typically emphasizes calibration, perhaps to the neglect of other, more important dimensions of external correspondence.

The purposes of this paper are twofold. First, a review is made of methods for analyzing various aspects of a particular quality of probabilistic forecasts—how well they actually anticipate what does and does not occur. The review focuses on the algebraic features of these techniques as well as how the procedures can be and are used in psychology and judgment analyses. The second purpose of the paper is to report the results of an empirical study of forecasting performance. The data of the study are examined in detail to provide a concrete illustration of one of the analytic techniques, that due to Yates (Note 1). The substantive aim of the study was to discover the relationship between forecasting behavior and self-judged knowledge of events.

INTERNAL CONSISTENCY VS EXTERNAL CORRESPONDENCE

Probabilistic forecasts are essentially statements by the forecaster of the degree to which he or she is certain that a particular event will occur

The assistance of Halimah Hassan and Bruce Carlson in the collecting and coding of data in the reported research is gratefully acknowledged. Requests for reprints should be sent to J. Frank Yates, 136 Perry Building, Department of Psychology, University of Michigan, Ann Arbor, MI 48104.

0030-5073/82/040132-25$02.00/0

some time in the future. Such forecasts are subject to the minimal constraint placed on probabilities that they be bounded by 0 and 1. Also, it is understood that the larger the reported probabilistic forecast is, the more certain the forecaster is that the event will indeed occur.

There is no guarantee that probabilistic forecasts will exhibit any of the other properties which are required to legitimately call them "probabilities." A collection of such judgments *are* appropriately considered to be probabilities when they do not violate commonly accepted axioms of probability theory, e.g., the Kolmogorov axioms (Apostol, 1962; Woodroofe, 1975). Then the judgments are said to be *internally consistent;* they are consistent with the axioms.

Internal consistency does not necessarily have anything whatsoever to do with what ultimately does or does not take place. It is entirely possible that a set of probabilistic forecasts will be perfectly internally consistent, yet be completely worthless in terms of anticipating what happens in the real world (Halmos, 1944). The extent to which probabilistic forecasts *do* anticipate the events at issue is called *external correspondence* (Yates, Note 1). While perfect internal consistency does not imply perfect external correspondence, perfect external correspondence trivially requires perfect internal consistency, of course.

It can be demonstrated (Ramsey, 1950; Winkler, 1972) that, if a person makes decisions on the basis of internally inconsistent probabilistic forecasts, that individual is vulnerable to cycles of transactions in which he or she is guaranteed to lose, regardless of which of the events being forecasted actually occurs. The extent to which internal inconsistency *does* lead to dysfunctional consequences in the real world is unknown. One does not have to stretch the imagination very far at all, however, to recognize the seriousness of deficiencies of forecast external correspondence. So, it is clear that the assessment, analysis, and understanding of external correspondence is a significant problem.

THE MEAN PROBABILITY SCORE

Conceptually, at least, the class of rules one might use to index the external correspondence of probabilistic forecasts is boundless. In practice, however, only a small number of such *scoring rules* are commonly used. Most of these rules are "proper," in that their structural properties are thought to discourage hedging when they are used as devices for rewarding the forecaster's performance, e.g., the logarithmic, spherical, and quadratic rules (Winkler & Murphy, 1968). Properness is not necessary, however, for a rule to be used for assessing the external correspondence of a collection of forecasts (Yates, Note 2).

By far the most widely employed rule for summarizing external correspondence is the *mean probability score* ($\overline{PS}$), also known as the *Brier*

score after Glenn W. Brier (1950), the meteorologist who introduced it. The mean probability score is a variant of the quadratic scoring rule. It can be described as follows.

Let the generic event in question, e.g., "Rain," "Dow Jones average rises," "Patient survives," be denoted by the letter A. Event A's occurrence or nonoccurrence is to be forecast on N different occasions. Denote by f_i the *probabilistic forecast* of event A's occurrence on the ith occasion, $i = 1, \ldots, N$. Also, define an *outcome index* for event A on the ith occasion by

$$\begin{aligned} d_i &= 1, \quad \text{if event } A \text{ occurs,} \\ &= 0, \quad \text{if event } A \text{ does not occur.} \end{aligned}$$

Then the *probability score* for occasion i is given by

$$\mathrm{PS}_i(f,d) = (f_i - d_i)^2. \tag{1}$$

Over all N different occasions, the *mean probability score* is then given by

$$\overline{\mathrm{PS}}(f,d) = \left(\frac{1}{N}\right)\sum_{i=1}^{N}(f_i - d_i)^2. \tag{2}$$

Clearly, $\overline{\mathrm{PS}}$ is 0 when external correspondence is perfect. $\overline{\mathrm{PS}}$ is 1 when forecasting performance is "counterperfect," i.e., $f_i = 0$ when event A occurs and $f_i = 1$ when event A does not occur.

THE SANDERS DECOMPOSITION OF $\overline{\mathrm{PS}}$

Sanders (1963) was apparently the first to offer a decomposition of $\overline{\mathrm{PS}}$ into components which index different aspects of external correspondence. The Sanders decomposition applies to forecasts that are restricted to a limited set of categories, e.g., tenths. Alternatively, the forecasts might be expressed continuously by the forecaster, but are then rounded or grouped into categories after the fact.

Given that forecasts are considered to be discrete, the method of computing $\overline{\mathrm{PS}}$ is slightly different in the situation treated by the Sanders decomposition as compared to the more general situation discussed previously:

$$\overline{\mathrm{PS}}(f,d) = \left(\frac{1}{N}\right)\sum_{j=1}^{J}\sum_{i=1}^{N_j}(f_j - d_{ij})^2, \tag{3}$$

where f_j is the jth allowable forecast, $j = 1, \ldots, J$, e.g., $f_4 = .3$ when forecasts are in tenths; d_{ij} is the outcome index for the ith occasion on which forecast f_j is offered; and N_j is the total number of occasions on which the forecast is f_j, $N = \sum_{j=1}^{J} N_j$.

Sanders (1963) shows that Eq. (3) can be expressed as

$$\overline{\mathrm{PS}}(f,d) = \left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j \bar{d}_j(1 - \bar{d}_j) + \left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j(f_j - \bar{d}_j)^2, \qquad (4)$$

where $\bar{d}_j = (1/N_j)\sum_{i=1}^{N_j} d_{ij}$ is the relative frequency of event A's occurrence over the N_j occasions when the forecaster reports forecast f_j. Equation (14) is what Yates (Note 1) calls the *Sanders decomposition* of $\overline{\mathrm{PS}}$.

Sanders (Note 3) referred to the first term on the right hand side of Eq. (4) as the "resolution" of the N forecasts. To avoid confusion with another statistic to be described below, Yates (Note 1) calls that term the *Sanders resolution* of the forecasts. Given that the Sanders resolution contributes to $\overline{\mathrm{PS}}$ positively, it is clear that the forecaster's goal should be to minimize the term. On the face of it, it appears the statistic is not under the forecaster's control, since it involves only outcome indexes, which are determined by the events. This is deceptive. Recall that occasions are sorted into classes $j = 1, \ldots, J$ according to the *forecaster's* reported forecasts for those occasions. To take an extreme example, if the forecaster always made the same prediction, say $f_j{}^*$, then $N_j{}^* = N$ and $N_j = 0$, for $j = 1, \ldots, J$ and $j \neq j^*$. This would mean that the Sanders resolution would be $\bar{d}(1 - \bar{d})$.

How can the forecaster minimize the Sanders resolution? The expression $(1/N)\sum_{j=1}^{J} N_j \bar{d}_j(1 - \bar{d}_j)$ achieves its smallest possible value of 0 when $N_j\bar{d}_j(1 - \bar{d}_j) = 0$ for each and every $j = 1, \ldots, J$. Now, $N_j\bar{d}_j(1 - \bar{d}_j) = 0$ only if $N_j = 0$, $\bar{d}_j = 1$, or $\bar{d}_j = 0$; $N_j = 0$ means that class j is not used. So, if class j *is* used, then either $\bar{d}_j = 1$ or $\bar{d}_j = 0$. If $\bar{d}_j = 1$, this means that event A occurs on *all* of the occasions for which forecast f_j is reported. If $\bar{d}_j = 0$, this means that event A occurs on *none* of the occasions for which forecast f_j is reported. The upshot of all this is that the forecaster minimizes the Sanders resolution if he or she never assigns the same forecast to two different occasions, one of which results in event A's occurrence, the other of which does not. A concrete example of ideal resolution, as absurd as it may seem, would be that in which all forecasting occasions that result in a rise in the Dow Jones average are assigned even-multiple forecasts by an analyst, e.g., .0, .2, . . . , while all other occasions are given odd-multiple forecasts.

The second expression on the right-hand side of Eq. (4), $(1/N)\sum_{j=1}^{J} N_j(f_j - \bar{d}_j)^2$, was called by Sanders (Note 3) the "reliability" of the set of N forecasts. Here, it is called the *reliability-in-the-small* to distinguish it from a similar, but different, component in the covariance decomposition of Yates (Note 1). Clearly, the forecaster's aim should be to minimize the reliability-in-the-small. This goal is achieved when, for each j, $j = 1, \ldots, J$, $f_j = \bar{d}_j$; i.e., the forecaster somehow manages to match individual discrete forecasts with their respective category relative frequencies. A given collection of forecasts is said to be *calibrated-in-*

the-small to the extent that this ideal of perfect matching is approached. Thus, for instance, a weather forecaster would exhibit good calibration-in-the-small if on 60% of the days when he forecasts a .6 chance of rain, it actually rains, and if on 30% of the days he forecasts a .3 chance of rain, rain occurs, and so forth.

THE MURPHY DECOMPOSITION OF $\overline{\text{PS}}$

Murphy (1972a, 1972b, 1973) has described several decompositions of $\overline{\text{PS}}$. His 1973 "new" decomposition is the $\overline{\text{PS}}$ decomposition that seems to be most widely used, at least in psychological research. Specialized to the type of situation presently under discussion involving forecasts for a single event A, Murphy's decomposition follows directly from Sanders'. It can be shown rather easily that the Sanders resolution can be expressed as

$$\left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j\bar{d}_j(1-\bar{d}_j) = \bar{d}(1-\bar{d}) - \left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j(\bar{d}_j - \bar{d})^2, \qquad (5)$$

where $\bar{d} = (1/N)\sum_{j=1}^{J}\sum_{i=1}^{N_j} d_{ij}$, the grand mean of the outcome index, is also the overall relative frequency of event A's occurrence. Substituting Eq. (5) into the Sanders decomposition, Eq. (4), we arrive at what Yates (Note 1) calls the *Murphy decomposition* of $\overline{\text{PS}}$:

$$\overline{\text{PS}}(f,d) = \bar{d}(1-\bar{d}) + \left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j(f_j - \bar{d}_j)^2 - \left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j(\bar{d}_j - \bar{d})^2. \quad (6)$$

Murphy (1973) has called the expression $(1/N)\sum_{j=1}^{J} N_j(\bar{d}_j - \bar{d})^2$ simply the "resolution" of the collection of forecasts. To distinguish this term from the Sanders resolution, Yates (Note 1) describes it as the *Murphy resolution*.

Bear in mind that $\bar{d}(1-\bar{d})$ is the variance of the outcome index. It is determined by "nature," or whatever it is that is responsible for event A's actual occurrence or nonoccurrence. So, effectively, what the Murphy decomposition does is divide the Sanders resolution into that part which is determined by outside forces (the outcome index variance) and that part which is controlled by the forecaster (the Murphy resolution). Given Eq. (5), however, it is apparent that the Murphy resolution is maximized under the same conditions that the Sanders resolution is minimized. So, both expressions reflect the same skill on the part of the forecaster, the ability to discriminate occasions when event A will and will not take place.

Analyses of probabilistic forecasting performance using the Sanders and Murphy decompositions are typically accompanied by "reliability diagrams" (Murphy & Winkler, 1977), in which relative frequencies $(\bar{d}_j)$

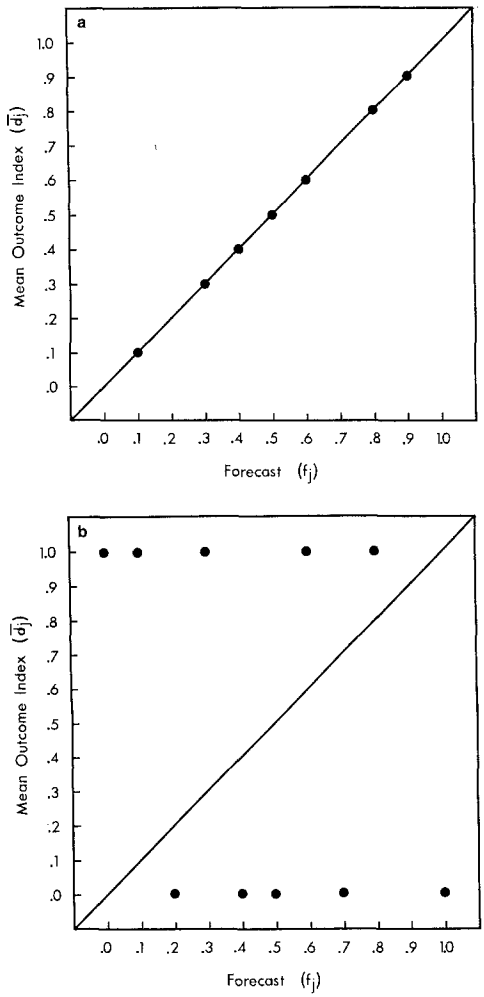

FIG. 1. Reliability diagram of (a) a hypothetical perfectly calibrated-in-the-small forecaster, (b) a hypothetical perfectly resolved forecaster, and (c) precipitation forecasts of Murphy and Winkler's (1977) Forecaster B (redrawn with the permission of the authors and the Royal Statistical Society).

are plotted against the respective discrete forecasts (f_j). When the points in a reliability diagram are connected to one another by lines, the resulting curve is often referred to as a "calibration curve" (Lichtenstein, Fischhoff, & Phillips, 1977).

Figures 1a–c are, respectively, reliability diagrams of the forecasts of a hypothetical perfectly calibrated-in-the-small forecaster, a hypothetical perfectly resolved forecaster, and Forecaster *B* in Murphy and Winkler's

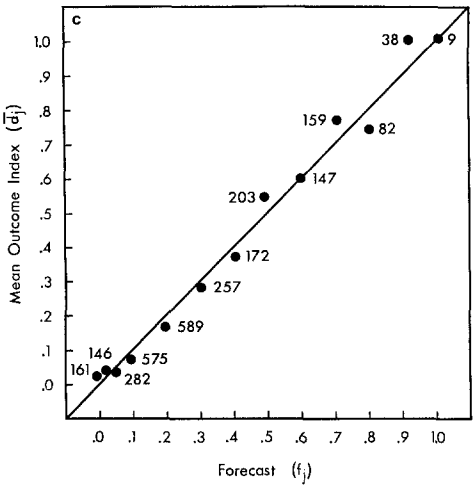

FIGURE 1 (*Continued*).

(1977) study of weather forecasters in Chicago. The event being forecasted by Forecaster *B* was precipitation 12 hr hence. Notice that Forecaster *B*'s performance was much closer to the ideal of perfect calibration-in-the-small than to that of perfect resolution. The reader should also recognize that Fig. 1b does not represent the only possible way that forecasts can be perfectly resolved. All that is required is that each point have a vertical coordinate of either zero or one.

COVARIANCE DECOMPOSITION OF $\overline{PS}$

Yates (Note 1) has derived what is called a covariance decomposition of $\overline{PS}$. In contrast to the Sanders and Murphy decompositions, the covariance decomposition can be applied to either continuous or discrete forecasts. The most basic form of the *covariance decomposition* of $\overline{PS}$ is given by

$$\overline{PS}(f,d) = S_d^2 + S_f^2 + (\bar{f} - \bar{d})^2 - 2S_{fd}, \tag{7}$$

where S_d^2 and S_f^2 are the variances of the outcome indexes and the forecasts, respectively, and S_{fd} is their covariance; $\bar{f}$ is, of course, the overall mean forecast for event *A*. Equation (7) is an instance of a well-known method of expressing a mean squared difference of two variables.

A perhaps more transparent and useful form of the covariance decomposition of $\overline{PS}$ is given by

$$\overline{PS}(f,d) = \bar{d}(1 - \bar{d}) + \Delta S_f^2 + S_{f,\min}^2 + (\bar{f} - \bar{d})^2 - 2S_{fd}, \tag{8}$$

where it is recognized (Yates, Note 1) that $S_{fd} = (\bar{f}_1 - \bar{f}_0)\bar{d}(1 - \bar{d})$, $S_{f,\min}^2 = (\bar{f}_1 - \bar{f}_0)^2\bar{d}(1 - \bar{d})$, and $\Delta S_f^2 = S_f^2 - S_{f,\min}^2$; $\bar{f}_1$ and $\bar{f}_0$ are, respectively, the

mean forecasts of event A's occurrence when event A does and does not actually occur. Several interpretative remarks and comments about Eq. (8) are in order.

The outcome index variance, $\bar{d}(1 - \bar{d})$, provides a good reference point for interpreting $\overline{PS}$. Suppose the forecaster always reports the constant forecast c, e.g., that there is always a 20% chance of rain. A careful examination of Eqs. (7) and (8) makes it clear that such a constant forecaster would achieve the following value of $\overline{PS}$:

$$\overline{PS}(c,d) = \bar{d}(1 - \bar{d}) + (c - \bar{d})^2. \tag{9}$$

From Eq. (9), it is clear that a constant forecaster can do no better than to report the relative frequency, i.e., set $c = \bar{d}$. Such relative frequency forecasters have sometimes been called "no skill" forecasters in the meteorological literature (Glahn & Jorgensen, 1970), seemingly because they do not make an attempt to do anything different from one occasion to the next. Perhaps, therefore, it is more appropriate to apply the adjective "no skill" to *any* constant forecaster. It takes considerable "baseline knowledge" to be able to set c near what $\bar{d}$ will eventually be. Consider, for instance, the task of forecasting defaults on personal loans. The average layperson is likely to have no idea of what the typical relative frequency of defaults is. Thus, if he or she were to attempt to forecast defaults with a constant probability c, that forecast would probably be much farther off the mark from $\bar{d}$ than a similar forecast offered by a banker experienced in the personal lending business.

Another reason that $\bar{d}(1 - \bar{d})$ is important as a reference point arises when one is interested in using $\overline{PS}$ as a means of comparing the skills of different forecasters. Again, a careful consideration of Eqs. (7) and (8) makes it clear that not only does $\bar{d}(1 - \bar{d})$ reflect aspects of forecasting performance *not* under the forecaster's control, but that the remaining terms in the decomposition index aspects that *are* under the forecaster's influence. Thus, it is the latter terms alone which should be used in making statements about relative forecasting abilities. Suppose, for instance, that two diagnosticians have considered two different large pools of cases. Diagnostician A achieves a value of $\overline{PS} = .13$, while Diagnostician B earns a score of $\overline{PS} = .23$, in anticipating whether the patients they considered did or did not have the disease in question. Clearly, the external correspondence of Diagnostician A was superior to that of Diagnostician B. If one were not careful, it would be tempting to conclude that Diagnostician A was more skilled. Suppose, however, it turned out that 45% of Diagnostician B's cases had the disease, while only 10% of Diagnostician A's did. A quick calculation then shows that the "skill" components of $\overline{PS}$ sum to .0400 for Diagnostician A and $-.0175$ for Diagnostician B. Thus,

we would come to the opposite conclusion regarding the forecasters' relative abilities.

The term $(\bar{f} - \bar{d})^2$ in the covariance decomposition is called by Yates (Note 1) the *reliability-in-the-large*. It indexes a performance characteristic labeled *calibration-in-the-large*. Like calibration-in-the-small, calibration-in-the-large reflects the ability of the forecaster to match mean forecasts to relative frequencies. In this instance, however, the matching applies to values over the entire collection of forecasts rather than to individual forecast categories. As implied in the discussion of $\bar{d}(1 - \bar{d})$, the ability of the forecaster to make $(\bar{f} - \bar{d})^2$ small might be seen as an indication of the quality of the forecaster's baseline knowledge about the event class under consideration. Alternatively, in some circumstances a large value of $(\bar{f} - \bar{d})^2$ might simply be a manifestation of a response bias.

While the reliability-in-the-large can often be interpreted as a measure of the forecaster's general knowledge about the event of interest, S_{fd}, the covariance of forecasts and outcome indexes, reflects the forecaster's ability to make distinctions among individual occasions on which the event might or might not take place. Thus, it could be thought of as assessing the sensitivity of the forecaster to specific signs which are indicative of what will happen in the future. It also shows whether that cue responsiveness is oriented in the proper direction. In a very real sense, the covariance indexes the heart of forecasting skill.

As suggested by Eqs. (7) and (8), the aim of the forecaster should be to minimize the variance of his or her forecasts, S_f^2. There is an obvious qualification on this advice, however. The only way S_f^2 can take on its absolute minimum possible value of zero is when the forecaster offers constant forecasts. This strategy would make the covariance term zero, too. So the proper objective of the forecaster should be to minimize S_f^2, *given* that he or she exercises his or her fundamental forecasting abilities, as represented by S_{fd}. The *conditional minimum forecast variance, given* S_{fd}, is $S_{f,\min}^2$. The conditional minimum value of S_f^2 is achieved under very interesting circumstances. $S_f^2 = S_{f,\min}^2$ when all forecasts for cases in which event A occurs are identical, i.e., $f_i = \bar{f}_1$, and all forecasts for cases in which event A does not occur are identical, i.e., $f_i = \bar{f}_0$. If, under these conditions, $\bar{f}_1 \neq \bar{f}_0$, one has a situation in which the forecaster has perfect foresight, in that he or she exhibits perfect discrimination of instances in which event A does and does not occur. The only thing that would possibly mar the forecaster's performance is mislabeling; the forecaster's numerical assignments would be inappropriate, i.e., $\bar{f}_1 < 1$ and $\bar{f}_0 > 0$. Since $\Delta S_f^2 = S_f^2 - S_{f,\min}^2$, ΔS_f^2 is appropriately considered as the "excess" variability in the given collection of forecasts. If S_{fd} indexes how responsive the forecaster is to information *related* to event A's occurrence, then

ΔS_f^2 might reasonably be taken as a reflection of how responsive the forecaster is to things that are *not* related to event A's occurrence.

Yates (Note 1) shows that there is a very straightforward relationship between the components of the covariance decomposition of $\overline{PS}$ and those of the decompositions of Sanders and Murphy. The essential observation demonstrated is a partitioning of the reliability-in-the-small:

$$\left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j(f_j - \bar{d}_j)^2 = S_f^2 + (\bar{f} - \bar{d})^2 - 2S_{fd} + \left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j(\bar{d}_j - \bar{d})^2. \quad (10)$$

Besides identifying how the covariance decomposition is related to the other decompositions, Eq. (10) shows that resolution and reliability-in-the-small are algebraically confounded with each other in a very direct way. A graphic summary of the relationships among the $\overline{PS}$ decompositions reviewed is presented in Fig. 2. As suggested in the figure, in a particular sense, the components of the covariance decomposition of $\overline{PS}$ are more basic than those of the other decompositions.

It might be noted that there exist forecasting situations in which three or more events form the relevant partition of the sample space, rather than simply event A and its complement, as assumed in the discussion to this point. Thus, for example, a political analyst might be required to offer a "vector forecast" of the probabilities that any one of three political parties, or none of them, will have the majority in the next parliament of a

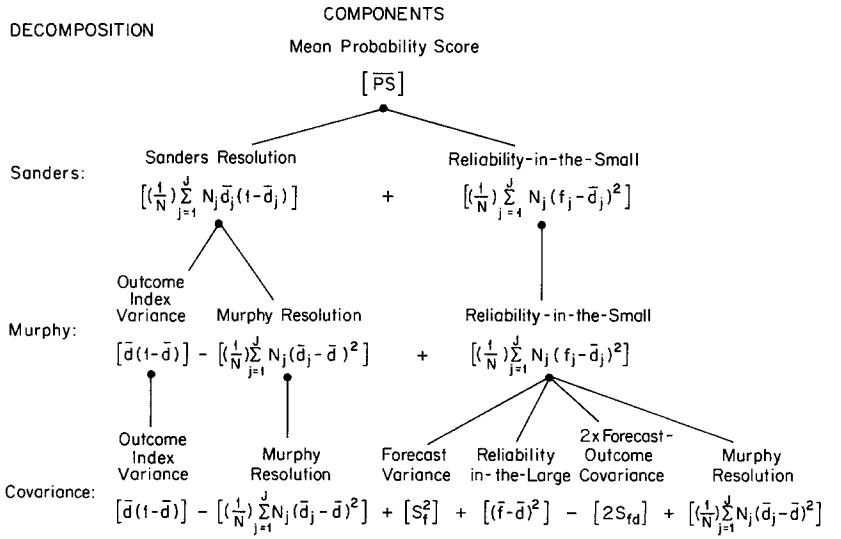

FIG. 2. Graphical display of relationships among the Sanders, Murphy, and covariance decompositions of the mean probability score.

certain country. There are vector versions of $\overline{PS}$ and the Murphy (1972a, 1972b, 1973) and covariance decompositions (Yates, Note 1) thereof. The reader is referred to the indicated papers for the details of those decompositions.

SELF-JUDGED KNOWLEDGE AND FORECASTING PERFORMANCE

A fundamental distinction made in the decision-making literature concerns the "level of uncertainty" surrounding the decision situation. An implicit continuum of uncertainty levels is often assumed to exist (cf. Coombs, Dawes, & Tversky, 1970, pp. 115–117). The continuum starts with "total ignorance" about which of the potential events relevant to a decision will occur. It then proceeds through the fuzzy kinds of uncertainty known as "ambiguity" (Ellsberg, 1961; Yates & Zukowski, 1976) and the more firmly established "risks," such as those associated with canonical events like the selection of colored marbles from an urn (Luce & Raiffa, 1957). The continuum ends, of course, with situations in which there is no uncertainty at all; at least in his or her own mind, the decision maker is absolutely sure of what the consequences of the given alternatives will be.

The substantive purpose of the study to be described presently was to find out how forecasters' self-judged knowledge about the events they are required to anticipate affects their reported forecasts for those events. Put another way, the aim was to determine the effect of level of uncertainty on forecasting behavior.

In the classical probability literature of the 17th and 18th centuries, perhaps the primary rule to be employed in assigning probabilities to events was the "principle of sufficient reason" (also known as the principle of *in*sufficient reason). Very often, the principle is associated with Laplace (1796/1951), who popularized it in his writings. In its essentials, the principle prescribes that, if there is no reason to think that one state of the world is any more likely to occur than any other, then the probabilities assigned to all states should be judged equal. It has been recognized for some time (e.g., Milnor, 1954) that there are some formal difficulties with this advice. Nevertheless, it is entirely possible that the same reasoning motivating the classicists could influence contemporary probabilistic forecasters, too. Thus, we might anticipate that forecasters would exhibit a tendency to "collapse" their forecasts for event A toward .5 when they feel that they know little about the conditions surrounding the event.

There is another plausible reason that forecasters might evidence the collapsing tendency described above. The forecaster could implicitly view the forecasting task as an exercise in estimating the outcome index by his or her forecast. The psychological loss function for the estimation procedure is quite conceivably single peaked (as is $\overline{PS}$, in form). This would be

the case if, say, the subject would feel regret or embarrassment about his or her forecast as an increasing function of its distance from the actual value of the index. For instance, one would probably feel pretty silly offering a forecast of .02 for rain and then observing a deluge.

Method

Subjects. Thirty-eight subjects volunteered to participate in the study. They were all individuals associated with the University of Michigan or Eastern Michigan University: 14 undergraduate students, 13 graduate students, and 11 faculty members. The subjects were not paid.

Task and instrument. The basic task requested of each subject was to submit probabilistic forecasts of the outcomes of several college basketball games that were to be played in various locations in the United States within 2 weeks of the time the subject made his or her predictions. It might be observed that there is nothing intrinsic to the sport of basketball that led to its selection as the experimental forecasting domain. The primary reasons basketball games were chosen for study were (a) at the time the investigation was conducted, basketball games were very numerous, (b) information about individual games seemed to vary in accessibility, and (c) subjects could be expected to differ from one another considerably in terms of how they might forecast outcomes.

The instructions and response scales required for the task were contained in a printed questionnaire which listed the home and visiting teams for 20 basketball games. For each game, the subject was asked to indicate which team he or she expected to win. Then the subject was requested to report how strongly he or she felt that the predicted winner would indeed win the game. This opinion was to be expressed in the form of a subjective probability from .5 to 1.0, corresponding to a slash the subject drew through a continuous, graded scale provided for the purpose. Finally, the subject was asked to rate his or her knowledge of the teams participating in the game. He or she indicated whether knowledge was "good," "fair," or "poor."

Subjects were allowed to keep the questionnaire for several days and to complete it at their leisure, as long as it was finished and returned to the investigator prior to the date on which the first game was played. Subjects were told explicitly that the questionnaire was to be completed without the assistance of others, since the investigator was interested in individual rather than group judgments.

Results

Event A for each of the 20 basketball games considered by the subjects was designated as "Home Team Wins," for the purposes of analysis. Thus, if for game i the subject predicted that the home team would win, f_i

was read directly from the subject's .5–1.0 subjective probability scale. If the subject expected the visiting team to win, f_i was taken to be 1 minus the reported subjective probability of the visiting team winning. Thus, complementarity of subjective probabilities for events A and $\bar{A}$ was imposed by fiat. Given the above specification of event A, $d_i = 1$ was assigned when the home team won in game i, and $d_i = 0$ was assigned when the visiting team won.

Illustrative covariance decompositions for individual forecasters. As an illustration of the use of the covariance decomposition of $\overline{PS}$ for individual forecasters, the responses of three subjects are examined in detail. The subjects in the study were ranked according to their mean probability scores over all 20 basketball games considered. Figure 3a–c displays the *covariance graphs* for the subjects with the best $\overline{PS}$, the $\overline{PS}$ that was one position better than the median, and the worst $\overline{PS}$, respectively. Obviously, a covariance graph is a close relative to the ordinary scatter plot. The abscissa is defined by outcome indexes, while the ordinate is defined by forecasts. Essentially, when the data are numerous enough, a covariance graph amounts to separate histograms for forecasts when event A does and does not take place, along with various summary statistics. Here, with fairly sparse data, histogram bars are replaced by rows of points for multiple forecasts. The horizontal and vertical dotted lines,

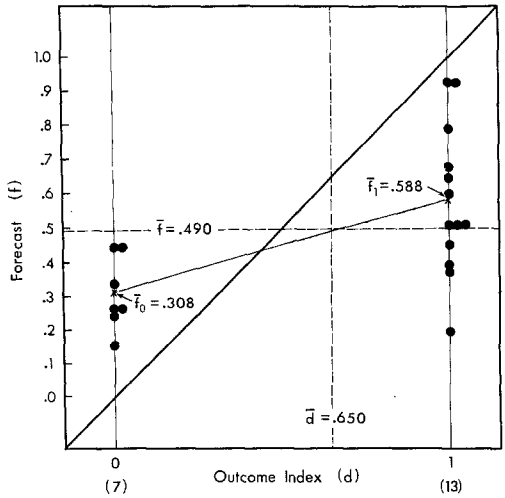

FIG. 3a. Covariance graph for the forecaster with the best value of $\overline{PS}$, Subject 3. $\overline{PS}$ decomposition:

$$\begin{aligned} \overline{PS} &= \bar{d}(1-\bar{d}) + \Delta S_f^2 + S_{f,\min}^2 + (\bar{f}-\bar{d})^2 - 2S_{fd} \\ [.1762] &= [.2275] + [.0328] + [.0506] + [.0254] - 2[.0637] \end{aligned}$$

$$\text{bias} = \bar{f} - \bar{d} = -.160, \qquad \bar{f}_1 - \bar{f}_0 = .280.$$

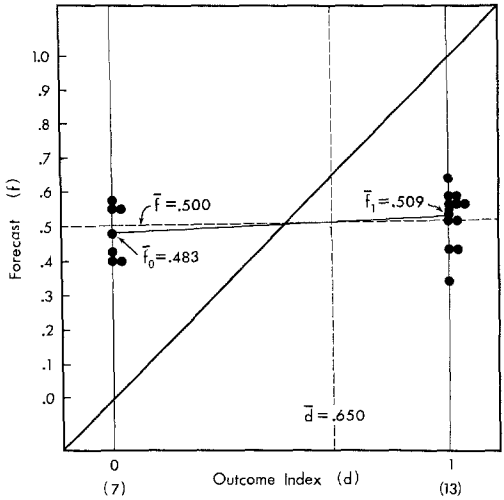

FIG. 3b. Covariance graph for the forecaster with the median+ value of $\overline{PS}$, Subject 16.

$\overline{PS}$ decomposition:

$$\begin{array}{rcl} \overline{PS} & = \bar{d}(1-\bar{d}) + \Delta s_f^2 + s_{f,\min}^2 + (\bar{f}-\bar{d})^2 - 2s_{fd} \\ [.2434] & = [.2275] + [.0052] + [.0002] + [.0225] - 2[.0060] \end{array}$$

$$\text{bias} = \bar{f} - \bar{d} = -.150, \quad \bar{f}_1 - \bar{f}_0 = .026.$$

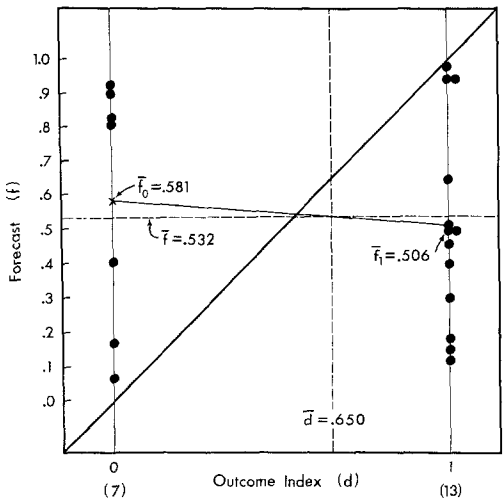

FIG. 3c. Covariance graph for the forecaster with the worst value of $\overline{PS}$, Subject 19.

$\overline{PS}$ decomposition:

$$\begin{array}{rcl} \overline{PS} & = \bar{d}(1-\bar{d}) + \Delta s_f^2 + s_{f,\min}^2 + (\bar{f}-\bar{d})^2 - 2s_{fd} \\ [.3686] & = [.2275] + [.0918] + [.0013] + [.0138] - 2[-.0171] \end{array}$$

$$\text{bias} = \bar{f} - \bar{d} = -.118, \quad \bar{f}_1 - \bar{f}_0 = -.075.$$

respectively, pass through the mean forecast and the mean outcome index, i.e., the overall relative frequency of event A. The heavy 45° line is referred to as the "veridical diagonal." The thin line connecting $(0,\bar{f}_0)$ and $(1,\bar{f}_1)$ is technically the regression line for forecasts on outcome indexes. As such, it and the $\bar{f}$ and $\bar{d}$ lines must all intersect at the point $(\bar{d},\bar{f})$.

Several distinguishing characteristics of the forecasting performance of the three subjects represented should be highlighted. First, of the three individuals, only the subject with the best $\overline{PS}$ did better than the relative frequency forecaster. All three subjects were rather poorly calibrated-in-the-large, offering forecasts that were generally biased far below the relative frequency with which home teams won. The forecasters with the best and worst values of $\overline{PS}$ exhibited a great deal of variability in their forecasts around $\bar{f}_0$ and $\bar{f}_1$, as evidenced by their covariance graphs and values of ΔS_f^2. By way of contrast, the median+ subject very seldom offered extreme forecasts of any kind. It is apparent, however, that the most important differences among the subjects pertain to the covariation of forecasts and outcome indexes, as represented by $\bar{f}_1 - \bar{f}_0$, since $\bar{d}(1 - \bar{d})$ is the same for all subjects. The subject with the best $\overline{PS}$ seemed to achieve his ranking primarily because of covariation; his forecast scatter was actually quite poor among the entire group of subjects. On the other hand, the median+ subject had a very minimal positive covariance. The modestness of his covariance was compensated for by the lack of scatter among his forecasts, however. The subject with the worst $\overline{PS}$ also had one of the worst covariances, a slightly negative value.

Covariance decomposition statistics over all subjects. The mean value of $\overline{PS}$ over all subjects was .2527. Only nine out of the 38 subjects, fewer than 25%, had values of $\overline{PS}$ better than .2275, the value for the relative frequency forecaster. Fifteen of the subjects, more than 39%, had values of $\overline{PS}$ worse than .2500, the value they would have achieved had they simply reported $f_i = .5$ for each game. So, on the whole, one would have to say that the subjects did not perform exceedingly well on the forecasting task.

Generally, subjects were very biased in their forecasts. The mean value of $\bar{f} - \bar{d}$ was $-.138$ ($p < .0001$). This also implies, of course, that the subjects were generally poorly calibrated-in-the-large.

Overall, the mean value of $S_{f,\min}^2$ was .0037. By comparison, the overall mean of the actual variances of subjects' forecasts, S_f^2, was .0379. Thus, on the average, the observed variability of subjects' forecasts was on the order of 10 to 11 times the variability that is necessary, given the covariance difference $\bar{f}_1 - \bar{f}_0$. In other words, subjects typically scattered their forecasts around $\bar{f}_0$ and $\bar{f}_1$ tremendously. This picture would look even worse if the two subjects who were constant forecasters were excluded from the analyses resulting in the above statistics.

The across-subject mean value of $\bar{f}_1 - \bar{f}_0$, the main ingredient of the covariance of forecasts and outcome indexes, was .074. While this value is statistically significant ($p < .0001$, under standard assumptions; larger realistically), its practical significance pales when one recalls that perfect forecasting performance would imply that $\bar{f}_1 - \bar{f}_0 = 1.0$. Perhaps a truer picture of just how poor the subjects generally were at anticipating game outcomes is gained from summarizing frequencies of covariance measures in particular categories. Two of the 38 subjects were constant forecasters, as indicated above; one always reported $f_i = .50$, while the other always reported $f_i = .60$. Seven of the remaining subjects had values of $\bar{f}_1 - \bar{f}_0$ that were negative, though none of these was statistically significantly different from zero. In all, only four subjects had values of $\bar{f}_1 - \bar{f}_0$ that *were* significantly different from zero ($p < .05$).

Self-judged knowledge relationships. Subjects' ratings of their knowledge (K) of the teams involved in each game were converted to numerical values according to the following scheme: good, 1; fair, 2; poor, 3. On the whole, the subjects felt that their knowledge of the teams playing in the games they considered was slightly less than fair ($\bar{K} = 2.238$, $SD = .535$). In part, this moderate degree of knowledgeability was built into the study intentionally. This goal was sought by selecting games that were distributed over the entire United States, although the largest concentration of games was in the eastern Midwest, the region surrounding the University of Michigan. In addition, care was taken to include games involving teams that are relatively obscure in the basketball world as well as games between traditional basketball "powerhouses."

The mean correlation r_{K-PS} between individual-game knowledge ratings and probability scores was $-.007$ (inverse of the mean of Fisher-transformed correlation coefficients). This mean was not statistically significantly different from zero (t test on Fisher-transformed correlation coefficients). Interestingly, only five of the subjects had values of r_{K-PS} that were statistically significant, and four of those were negative. The apparent lack of association between self-reported team knowledge and forecasting performance level manifested itself at the aggregate level, too. The correlation between $\bar{K}$ and $\overline{PS}$ was $r_{\bar{K}-\overline{PS}} = .058$ (ns).

The statistic $E_i = (f_i - .5)^2$ was defined as an index of the extremeness of each forecast from .5. The mean value of r_{K-E} across subjects was $-.648$ (inverse of the mean of Fisher-transformed correlation coefficients). This value is highly significantly different from zero ($p < .0001$, via t test on Fisher-transformed correlation coefficients). Thus, it is very clear that when subjects felt that they knew little about the teams in a given game, they were very much inclined to offer a middle-of-the-road forecast near .5. That is, the "collapsing" tendency hypothesis was confirmed.

Discussion

The results of the study pose an interpretation problem for probabilistic forecasts. A forecast near .5 might mean that the forecaster is privy to a great deal of conflicting evidence which thus forces the forecaster to firmly judge the odds of the given event's occurrence to be just about even. Alternatively, such a report might mean that the forecaster simply does not feel that he or she knows very much about the situation. Thus, it would seem that a sensible thing to request of forecasters is not only their forecasts, but also an indication of how secure they feel about their knowledge of the given conditions. Unfortunately, however, the present study suggests that such self-judged knowledge is unlikely to be predictive of the accuracy of the forecaster's opinions. The issues obviously beg for further study, at both the fundamental and prescriptive levels. An initial concern of such research is the generalizability of the present results. The subjects in the current research were amateurs. It is possible that well-trained, professional forecasters would, in fact, be able to tell reliably when they do and do not know things that are useful to their task.

USES OF $\overline{PS}$ DECOMPOSITIONS

There are numerous research and practical situations in which decompositions of $\overline{PS}$ can be and have been fruitfully employed. This section is devoted to a survey of these applications.

General Forecasting Performance Analyses

Seemingly, there are endless circumstances in which people make probabilistic judgments. It is only natural that there is considerable interest in knowing just how good those judgments typically are, in terms of external correspondence. Beyond the question of how good such probabilistic judgments are in a general sense, the issue of how good they are in very specific ways is of interest, too. And that is where the usefulness of $\overline{PS}$ decompositions becomes most salient.

In recent years there have been an increasing number of studies in which investigators have reported on particular aspects of external correspondence. A most curious feature of this trend of research has been an almost exclusive emphasis on calibration-in-the-small, to the neglect of other performance aspects. Thus, for instance, in their discussion of the quality of subjective probability distributions for various random variables, Tversky and Kahneman (1974) write only of "calibration" (calibration-in-the-small). Rather than demonstrate its significance, Tversky and Kahneman simply assume that calibration-in-the-small is a desirable quality for a judge to pursue, considering a particular type of miscalibration to be a "bias . . . common to naive and to sophisticated

subjects'' (p. 1129). They also point out that the bias ''is not eliminated by introducing proper scoring rules, which provide incentives for external calibration'' (p. 1129). What Tversky and Kahneman do not acknowledge is that proper scoring rules provide incentives for *all* aspects of external correspondence, not just calibration.

It often appears that the general construct of external correspondence itself is superseded in significance in the literature by calibration-in-the-small. For example, Lichtenstein, Fischhoff, and Phillips (1977) devote an entire review to calibration-in-the-small, barely mentioning other external correspondence dimensions. Lichtenstein and Fischhoff (1980) have reported an elaborate experimental training program, the stated purpose of which was to enhance judges' calibration-in-the-small. Fryback and Erdman (1979) have urged similar training efforts to improve the calibration-in-the-small of physicians' probabilistic opinions. Christensen-Szalanski and Bushyhead's (1980) analysis of physicians' probabilistic diagnoses in actual clinical settings focused almost entirely on calibration-in-the-small.

An interesting exception to this seemingly complete emphasis on calibration is represented by the paper by Shapiro (1977). Although he uses slightly different language and expressions, in so many words, Shapiro recommends that the logarithmic scoring rule be used as a research tool in assessing the external correspondence of physicians' probabilistic judgments. The mean logarithmic score over N judgments can be expressed as

$$\overline{\mathrm{Lg}}(f,d) = \left(\frac{1}{N}\right)\sum_{i=1}^{N} \log\left[f_i d_i + (1-f_i)(1-d_i)\right], \tag{11}$$

where the symbols have the same meanings as before. It is worth noting that $\overline{\mathrm{PS}}$ is generally to be preferred to $\overline{\mathrm{Lg}}$ as an index of external correspondence, since a single instance in which the subject indicates $f_i = 1$ where $d_i = 0$ or $f_i = 0$ when $d_i = 1$ makes $\overline{\mathrm{Lg}} = -\infty$, regardless of what happens on all other occasions. In any case, Shapiro indicates a decomposition of $\overline{\mathrm{Lg}}$ into components that roughly index aspects of performance related to calibration-in-the-large and other performance characteristics. Specifically, he suggests computing $\overline{\mathrm{Lg}}(\bar{f},d)$, the mean value of Lg one would achieve by always forecasting the physician's mean forecast $\bar{f}$. Shapiro refers to $\overline{\mathrm{Lg}}(\bar{f},d)$ as the score due to the physician's ''anchor point'' $\bar{f}$. He speculates that, in the medical context, ''a correct anchor-point probability may be obtained either through knowledge of the literature or by extensive clinical experience'' (p. 1512). Shapiro indicates that the difference $\overline{\mathrm{Lg}}(f,d) - \overline{\mathrm{Lg}}(\bar{f},d)$ can be taken as an index of the physician's ability ''to individualize assessments to the unique characteristics of the patient'' (p. 1512). Shapiro (1977, p. 1513) implies, but does not

prove, that $\overline{\text{Lg}}(f,d) - \overline{\text{Lg}}(\bar{f},d)$ can be shown to reflect directly such individualization of predictions. As we have seen, it is already known that decompositions of $\overline{\text{PS}}$ achieve the goal of partitioning important performance aspects even more finely than does Shapiro's technique.

From the perspective provided by the Sanders decomposition of $\overline{\text{PS}}$, the aspect of probabilistic judgment external correspondence which is implicitly neglected by the current overemphasis on calibration-in-the-small is resolution. Recall that, according to the Sanders decomposition, $\overline{\text{PS}}$ is partitioned into reliability-in-the-small, which indexes calibration-in-the-small, and the Sanders resolution. Despite their exclusive focus on the calibration-in-the-small of physicians' probabilistic judgments, Fryback and Erdman (1979) recognize that "even if calibration improves, it will remain to show this benefits patients" (p. 344). Sanders (1973), summarizing years of research on probabilistic weather forecasting, indicates that the reliability-in-the-small typically accounts for a much smaller share of $\overline{\text{PS}}$ than does the Sanders resolution. He goes on to suggest that forecasting "skill would not be significantly advanced if bias [miscalibration] were entirely absent" (p. 1176). He shows that the deterioration of the quality of weather forecasting performance over the length of the forecasting horizon (1 day vs 4 days) is largely due to the deterioration of resolution. Thus, it seems that, in at least some domains, the preoccupation with calibration-in-the-small is perhaps unwarranted.

Why is calibration-in-the-small considered so important and resolution ignored? Part of the explanation seems to be that calibration-in-the-small appears so intuitively reasonable—on the surface. A reliability diagram with points close to the 45° diagonal just "looks right." When a stock price rises on 20% of the occasions when Market Analyst A says there is a 20% chance it will rise, and 70% of the time when he says there is a 70% chance it will rise, that feels right, too. Suppose, however, that prices rise on 100% of the occasions when Market Analyst B says there is a 20% chance of a price rise, and they never rise when he says there is a 70% chance of a price rise. Assuming that only forecasts of 20 and 70% are considered, Analyst A is perfectly calibrated-in-the-small, while Analyst B is perfectly resolved. A smart investor should prefer the services of Analyst B. He could make a fortune by selling stocks when Analyst B offers a 70% chance that their prices will rise and buying them when Analyst B says there is a 20% chance that their prices will rise. So, while the appeal of calibration-in-the-small is largely "aesthetic," the practical significance of resolution can potentially be much greater. In essence, resolution pertains to a much more fundamental skill than calibration; it refers to the ability of the forecaster to discriminate individual occasions on which the event of interest will and will not take place. By contrast, calibration concerns the forecaster's ability to assign the "right" numeri-

cal labels to his or her forecasts. Of course, such proper labels do permit one to interpret probabilistic forecasts the same as relative frequency probability estimates.

Another part of the reason that calibration-in-the-small seems to have taken precedence over resolution is that resolution measures are sometimes difficult to understand and have, in fact, often been misunderstood. For instance, Lichtenstein, Fischhoff and Phillips (1977) incorrectly report that the Murphy resolution "reflects the ability of the assessor to sort the events into subcategories for which the hit rate is maximally different from the overall hit rate." As we have seen, the Murphy resolution actually reflects the ability of the assessor to sort events into categories for which the hit rates are either 0 or 1. Lichtenstein and Fischhoff (1977) indicate that the Murphy resolution "measures the ability of the responder to discriminate different degrees of subjective uncertainty" (p. 162). The truth of the matter is that one could compute and interpret the Murphy resolution even if the response categories were completely nonnumerical. In the same article, Lichtenstein and Fischhoff (1977, p. 175) imply that one can judge resolution from the slopes of calibration curves. Figure 1b illustrates the fact that resolution and calibration curve slopes have no necessary relationship to each other at all. Lichtenstein and Fischhoff (1977, p. 162) also indicate that resolution and calibration are independent of each other. As shown by Eq. (10) and Fig. 2, this is clearly not the case. The expression for the Murphy resolution, $(1/N)\sum_{j=1}^{J} N_j(\bar{d}_j - \bar{d})^2$, is genuinely more difficult to interpret intuitively than is the expression for the Sanders resolution, $(1/N)\sum_{j=1}^{J} N_j\bar{d}_j(1 - \bar{d}_j)$. Thus, the misunderstandings and lack of appreciation for resolution are not surprising.

Study of Forecasting Processes

Decompositions of $\overline{PS}$ hold the potential of being valuable tools in the study of basic judgment processes. Specifically, one can construct reasonable arguments as to the foundations of variations in the various aspects of judgment performance indexed by decomposition components. One can then design experiments to test for the effects of manipulations of the hypothesized underlying factors. The experiment reported by Mehle, Gettys, Manning, Baca, and Fisher (Note 4) might be thought of in this light. In that study, the authors tested the effects of auxiliary tasks on subjects' tendencies to offer excessively low probabilities for unspecified "catch-all" hypotheses. The authors compared their subject groups' performance with respect to the components of the vector version of the Murphy decomposition of $\overline{PS}$. Unfortunately, the conclusions one can draw from the analyses are limited, due to the fact that Mehle *et al.* did not compute $\overline{PS}$ and the component scores in the standard way, defining

the outcome index as an indicator variable. Instead, the outcome index was taken to be a population relative frequency.

An alternative strategy one could take in using $\overline{PS}$ decompositions to study basic judgment processes might proceed in the opposite direction. That is, one might observe natural covariation between $\overline{PS}$ decomposition components and other variables and then pursue what seem to be plausible explanations for that observed covariation. For instance, Sanders' (1973) finding that the resolution of weather forecasts, but not their calibration, diminished as a function of the forecast horizon suggests that the fundamental ability of the forecaster to tell what will and will not happen is what suffers over time. Perhaps a more fine-grained analysis using the covariance decomposition of $\overline{PS}$ might provide even more specific guidance as to appropriate hunches to entertain as explanations for the forecast horizon effect.

Forecaster Evaluation and Selection

It has already been indicated how $\overline{PS}$ decompositions can be used to evaluate the comparative abilities of forecasters. Going one step further, one could use the results of such decomposition analyses to more appropriately reward forecasting performance and to select skillful forecasters. The assumptions implicit in these suggestions are that the samples of forecasting occasions on which $\overline{PS}$ measures are computed are representative of the forecasting situations of interest and that the samples are large enough to justify the belief that the values of the relevant statistics provide good estimates of their population counterparts. But, what is "large enough?"

It has often been assumed that the sampling distributions of the components of the Murphy decomposition are especially formidable (Lichtenstein *et al.*, 1977). The simplicity of the covariance decomposition of $\overline{PS}$ suggests that the sampling distributions of such components probably are not as peculiar as originally feared. Preliminary simulation results in our laboratory (to be reported in another article) suggest that this is indeed the case. Thus, the prospect of being able to make reliable judgments about forecasters' skills on the basis of fairly small samples of their products seems promising.

Forecaster Training

From a practical standpoint, the most enticing possibility offered by $\overline{PS}$ decompositions is their use in forecaster training. Most schemes for the training of forecasters have had feedback as the primary training device. Staël von Holstein (1972) provided his expert and nonexpert subjects with feedback on their performance in predicting stock market activity. The

feedback was primarily in the form of the subjects' scores computed via the quadratic scoring rule. The effectiveness of this feedback was generally disappointing. Staël von Holstein (1972, p. 144) acknowledged that "it is possible that there was too much information in the feedback for some (or all) participants to assimilate." Because the forecaster could possibly focus his or her attention on only one performance dimension at a time, it seems that feedback on $\overline{PS}$ decomposition components holds considerably more promise as a means of improving forecasting performance than Staël von Holstein's global approach.

Lichtenstein and Fischhoff (1980) attempted such a component-wise feedback training procedure. Using primarily factual or "almanac" questions, Lichtenstein and Fischhoff sought to demonstrate the effectiveness of providing feedback on components of the Murphy decomposition for improving their subjects' performance. The results showed that reliability-in-the-small scores generally did, in fact, improve, while Murphy resolution scores remained essentially the same. This pattern of results might have been due to the fact that, while subjects were provided with detailed instruction in how to interpret reliability-in-the-small scores, they were apparently told little or nothing about how to interpret resolution scores. It would be of considerable interest to know exactly how Lichtenstein and Fischhoff's subjects improved their calibration-in-the-small. As implied by Eq. (10) and Fig. 2, such improvement might have been achieved by nothing more than an uninteresting general translation of judgments, leading to a reduction in miscalibration-in-the-large. On the other hand, the improvement might have been produced by a reduction in judgment variance or an increase in judgment–outcome index covariation. These would be impressive accomplishments indeed.

There are good reasons for being cautious about the significance of feedback training in general and feedback studies involving factual questions in particular for the improvement of forecasting skill. Since Alpert and Raiffa (Note 5) used them in an early study on Harvard Business School students' probabilistic judgment tendencies, researchers have used almanac questions in a wide variety of investigations. These studies have revealed a number of phenomena that are truly interesting in their own right. It should be recognized, however, that there are some essential differences between true forecasting situations and the circumstance of answering almanac questions. Perhaps the most important difference is that one can answer an almanac question definitively by consulting an almanac, whereas in a true forecasting situation, because the event in question has not already occurred, no mortal being can know definitively whether or not the event will occur. Effectively, probabilistic judgments about answers to almanac questions are statements about the perception of one's own state of knowledge. By way of contrast, true forecasts are

statements about not only the forecaster's knowledge, but also possibly the "inherent" accessibility of predictive information.

As implied in the previous discussion, from the perspective provided by the Sanders decomposition of $\overline{PS}$, a primary aim one should have in a training program is the improvement of resolution. It is informative to think about what perfect resolution of probabilistic judgments about factual questions would mean. To be concrete, think of the illustrative question of Lichtenstein and Fischhoff (1980, p. 151), "which is longer, the Suez Canal or the Panama Canal?" The subject must indicate which possible answer he or she thinks is correct, "Suez" or "Panama," and then report a probability between .5 and 1.0 that the stated answer is indeed correct. In the coding of such responses, the generic event A is defined as "my indicated answer is correct." The outcome index d is 1 if the indicated answer *is* correct; it is 0 otherwise. Now, suppose that resolution were perfect. Thus, each summand of the Sanders resolution, $(1/N)\sum_{j=1}^{J} N_j \bar{d}_j(1 - \bar{d}_j)$, would be zero. What this would mean is that the subject is capable of perfectly discriminating those occasions when his or her answers are correct from those occasions when those answers are not correct. It is hard to imagine how a person could do this without knowing the correct answers themselves. The point is that the interpretation of $\overline{PS}$ decomposition components for factual questions is peculiar and of limited value for understanding true forecasting performance.

Given that resolution and the variance and covariance components of the covariance decomposition of $\overline{PS}$ concern aspects of forecasting behavior which go beyond simply the way the forecaster assigns numbers, it does not appear that they can be materially affected by mere feedback alone. Instead, it seems that to influence these terms, one must induce the forecaster to use different predictive information than he or she ordinarily uses or to employ such information in a different way than is customary. That is, a fundamentally different approach to forecaster training must be taken.

REFERENCES

Apostol, T. M. *Calculus, Vol. 2.* New York: Blaisdell, 1962.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review,* 1950, **78**(1), 1–3.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance,* 1981, **7**, 928–935.

Coombs, C. H., Dawes, R. M., & Tversky, A. *Mathematical psychology.* Englewood Cliffs, N.J.: Prentice–Hall, 1970.

Ellsberg, D. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics,* 1961, **75**, 643–649.

Fryback, D. G., & Erdman, H. Prospects for calibrating physicians' probabilistic judgments: Design of a feedback system. *IEEE Proceedings,* 1979, 340–344.

Glahn, H. F., & Jorgensen, D. L. Climatological aspects of the Brier *P*-score. *Monthly Weather Review,* 1970, **98,** 136–141.

Halmos, P. R. The foundations of probability. *American Mathematical Monthly,* 1944, **51,** 493–510.

Laplace, P. S. *A philosophical essay on probabilities.* New York: Dover, 1951. (Originally published in French in 1796.)

Lichtenstein, S., & Fischhoff, B. Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance,* 1977, **20,** 159–183.

Lichtenstein, S., & Fischhoff, B. Training for calibration. *Organizational Behavior and Human Performance,* 1980, **26,** 149–171.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. Calibration of probabilities: The state of the art. In H. Jungermann & G. deZeeuw (Eds.), *Decision making and change in human affairs.* Dordrecht, Holland: Riedel, 1977. Pp. 275–324.

Luce, R. D., & Raiffa, H. *Games and decisions.* New York: Wiley, 1957.

Milnor, J. Games against nature. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision processes.* New York: Wiley, 1954.

Murphy, A. H. Scalar and vector partitions of the probability score: Part I. Two-state situation. *Journal of Applied Meteorology,* 1972, **11,** 273–282. (a)

Murphy, A. H. Scalar and vector partitions of the probability score: Part II. *N*-state situation. *Journal of Applied Meteorology,* 1972, **11,** 1183–1192. (b)

Murphy, A. H. A new vector partition of the probability score. *Journal of Applied Meteorology,* 1973, **12,** 595–600.

Murphy, A. H., & Winkler, R. L. Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics,* 1977, **26,** 41–47.

Ramsey, F. P. *The foundations of mathematics and other logical essays.* New York: Humanities Press, 1950.

Sanders, F. On subjective probability forecasting. *Journal of Applied Meteorology,* 1963, **2,** 191–201.

Sanders, F. Skill in forecasting daily temperature and precipitation: Some experimental results. *Bulletin of the American Meteorological Society,* 1973, **54,** 1171–1179.

Shapiro, A. R. The evaluation of clinical prediction. *New England Journal of Medicine,* 1977, **296,** 1509–1514.

Staël von Holstein, C.-A. S. Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance,* 1972, **8,** 139–158.

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

Winkler, R. L. *Introduction to Bayesian inference and decision.* New York: Holt, Rinehart & Winston, 1972.

Winkler, R. L., & Murphy, A. H. "Good" probability assessors. *Journal of Applied Meteorology,* 1968, **7,** 751–758.

Woodroofe, M. *Probability with applications.* New York: McGraw–Hill, 1975.

Yates, J. F., & Zukowski, L. G. Characterization of ambiguity in decision making. *Behavioral Science,* 1976, **21,** 19–25.

REFERENCE NOTES

1. Yates, J. F. *Forecasting performance: a covariance decomposition of the mean probability score.* Paper presented at the 22nd Annual Meeting of the Psychonomic Society, Philadelphia, November 1981; also an unpublished manuscript.
2. Yates, J. F. *Scoring rules for forecasts* (Tech. Rep. No. 16) Michigan–Chicago Cognitive Science Program, April 1981.

3. Sanders, F. *The evaluation of subjective probability forecasts* (Tech. Rep. No. 5, Contract AF 19(604)-1305) Department of Meteorology, Massachusetts Institute of Technology, 1958.
4. Mehle, T., Gettys, C., Manning, C., Baca, S., & Fisher, S. *The availability explanation of excessive plausibility assessments.* Decision Processes Laboratory Report No. TR 30-7-79, Norman: Univ. of Oklahoma, July 1979.
5. Alpert, M., & Raiffa, H. *A progress report on the training of probability assessors.* Unpublished manuscript, Harvard Univ., 1969.

RECEIVED: June 25, 1981