ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 41, 281–299 (1988)

Analyzing the Accuracy of Probability Judgments for Multiple Events: An Extension of the Covariance Decomposition

J. FRANK YATES

The University of Michigan

The probability score (PS) can be used to measure the overall accuracy of probability judgments for a single event, e.g., "Rain falls," or "This patient has cancer." It has been shown previously how a "covariance decomposition" of the mean of PS over many occasions indexes several distinct aspects of judgment performance (J. F. Yates, *Organizational Behavior and Human Performance*, **30**, 132–156 (1982)). There are many situations in which probability judgments are reported for sample space partitions containing more than one event and its complement, e.g., medical situations in which a patient might suffer from Disease X, Disease Y, or Disease Z, or testing situations in which the correct answer to an item might be any one of alternatives (a) through (e). The probability score for multiple events (PSM) serves as a measure of the overall accuracy of probability judgments for the events in partitions of any size. The present article describes and interprets an extension of the covariance decomposition to the mean of PSM. The decomposition is illustrated with data from two contexts, medicine and education.

Imagine a labor consultant who considers how a union might respond to an offer from management. Four events are recognized as possibilities: A_1 = "Accept offer," A_2 = "Reject offer, but continue negotiating," A_3 = "Reject offer, then strike," and A_4 = "Other actions," e.g., "Reject offer, then slow down." The expert *could* make a deterministic judgment that one of these events definitely will occur. However, the consultant accepts the arguments for probabilistic assessments, e.g., the fact that they allow the decision maker to trade off uncertainty against value in, say, decision analyses. So the following probability judgments are reported: $P'(A_1) = .60$, $P'(A_2) = P'(A_3) = .15$, $P'(A_4) = .10$.

The present article concerns the following issue: How can we characterize and analyze the accuracy with which probability judgments for K-event partitions of a sample space ($K \geq 2$) anticipate the events that ultimately occur? For instance, over many different labor disputes, just how good are the labor consultant's predictions; i.e., precisely how expert *is*

It is my great pleasure to acknowledge Keith Levi's and Shawn Curley's critical reading of an earlier version of this paper. I also appreciate Halimah Hassan's assistance with data analyses. Requests for reprints should be sent to J. Frank Yates, 136 Perry Building, Department of Psychology, University of Michigan, Ann Arbor, MI 48104.

0749-5978/88 $3.00

the expert? How does that individual's performance compare with that of other authorities? It is well known that there are important and reasonably distinct components of the overall accuracy of probability judgments for single events (cf. Yates, 1982), e.g., calibration, resolution, bias, and slope. What are significant aspects of accuracy for multiple-event judgments?

The *probability score* (PS), also known as the *Brier score*, is the most commonly used measure of probability judgment accuracy. Virtually all psychological research involving PS has employed its single-event form. However, even in his earliest papers on PS, Brier (1950) indicated that PS can be applied to multiple-event partitions, too. The concepts of reliability, i.e., "calibration," as it is most often described in psychology, and resolution are now common in discussions of single-event probability judgment accuracy. Sanders (1963) and Murphy (1972a, 1973) have shown how $\overline{PS}$, the mean of PS over several single-event judgment occasions, can be decomposed into various components, including measures for calibration and resolution. Yates (1982) has described a "covariance decomposition" which parcels $\overline{PS}$ into terms reflecting other aspects of single-event judgment accuracy.

Murphy (1972b) has demonstrated how his single-event $\overline{PS}$ decomposition can be extended to the multiple-event case. The purpose of the present paper is to describe and illustrate a similar extension of the covariance decomposition of $\overline{PS}$ to multiple-event situations. It is surprising that, despite the existence for some time of multiple-event methods such as Murphy's $\overline{PS}$ decomposition, almost all analyses of probability judgment accuracy in the psychological literature have considered only the single-event case. In the above labor consultant example, single-event methods would allow us to examine in detail how the expert makes judgments for, say, the event "Accept offer." But they would permit us to say nothing about all the other possibilities. As will be shown, it is often the case that considerable insights *can* be gained about judgment variation over several events.

REVIEW OF THE SINGLE-EVENT COVARIANCE DECOMPOSITION

As a review, consider the case of a single target event A, e.g., "Rain," to be concrete.[1] The *probability judgment* for the target event is denoted $f = P'(A)$ and is, of course, between 0 and 1. The *outcome index* on a given occasion is defined as

$$d = 1, \quad \text{if event } A \text{ occurs}$$

[1] More complete discussions of single-event decompositions of $\overline{PS}$ can be found in articles by Yates (1982, 1984) and Yates and Curley (1985).

$$= 0, \quad \text{if event } A \text{ does not occur.} \tag{1}$$

The outcome index can be seen as the probability judgment of a clairvoyant, "God's probability," as some put it. The probability score is formally defined as the squared difference between f and d:

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

Clearly, $0 \leq \mathrm{PS} \leq 1$, and a judge's objective should be to minimize PS. Over N occasions, indexed by $i = 1, \ldots, n$, the mean of PS is given by

$$\overline{\mathrm{PS}}(f,d) = (1/N) \sum_{i=1}^{N} (f_i - d_i)^2. \tag{3}$$

The $\overline{\mathrm{PS}}$ *covariance decomposition* (Yates, 1982; Yates & Curley, 1985) can be expressed as

$$\begin{aligned} \overline{\mathrm{PS}}(f,d) = {} & \mathrm{Var}(d) + \mathrm{MinVar}(f) + \mathrm{Scat}(f) \\ & + \mathrm{Bias}^2 - 2[\mathrm{Slope}][\mathrm{Var}(d)]. \end{aligned} \tag{4}$$

The various components of $\overline{\mathrm{PS}}$ on the right-hand side of Eq. (4) have the following definitions and interpretations:

Var(d) is the variance of the outcome index, which, in its simplest form, can be written as

$$\mathrm{Var}(d) = \bar{d}(1 - \bar{d}), \tag{5}$$

where

$$\bar{d} = (1/N) \sum_{i=1}^{N} d_i \tag{6}$$

is the relative frequency, or "base rate," with which the target event A occurs. In many situations, the target event is completely outside the judge's control, e.g., when A = "Rain" or A = "IBM stock declines." In such instances, d and, hence, the Var(d) part of $\overline{\mathrm{PS}}$ are beyond the judge's influence, too.

The bias statistic is defined by

$$\mathrm{Bias} = \bar{f} - \bar{d}, \tag{7}$$

in which

$$\bar{f} = (1/N) \sum_{i=1}^{N} f_i \tag{8}$$

is the mean probability judgment reported for the target event. The bias statistic reflects a type of overall miscalibration, the extent to which the judge's assessments are generally too high or too low, i.e., by how much they are biased. The square of the bias, which is what actually appears in the covariance decomposition, is sometimes called the "reliability-in-the-large" (RIL), i.e.,

$$\text{RIL} = \text{Bias}^2. \tag{9}$$

Clearly, RIL characterizes overall miscalibration irrespective of the direction of the error.

The term Slope in Eq. (4) is given by

$$\text{Slope} = \bar{f}_1 - \bar{f}_0, \tag{10}$$

where

$$\bar{f}_1 = (1/N_1) \sum_{j=1}^{N_1} f_{1j} \tag{11}$$

is the conditional mean probability judgment for the target event A over those N_1 occasions when that event actually occurs (hence the f_{1j} notation, corresponding to $d = 1$); $\bar{f}_0$ is defined similarly for the remaining N_0 instances when the target event does not occur, with $N = N_1 + N_0$. In the ideal case, the judge always reports $f = 1$ when the target event is going to occur and $f = 0$ when it is not. This situation would yield the maximum possible value of Slope = 1. Thus, it makes sense for slope to contribute to $\overline{\text{PS}}$ negatively, as it does. The name for the statistic is appropriate because it is literally the slope of the regression line when probability judgments are regressed on outcome indexes. The term "covariance decomposition" applies because slope is so important to $\overline{\text{PS}}$ and because the covariance of the judge's assessments and the outcome indexes can be expressed as

$$\text{Cov}(f,d) = [\text{Slope}][\text{Var}(d)], \tag{12}$$

which, with a multiplier of 2, is the last term in the decomposition.

The definition of the Scat(f) term in the covariance decomposition of $\overline{\text{PS}}$ is given by

$$\text{Scat}(f) = (1/N)[N_1\text{Var}(f_1) + N_0\text{Var}(f_0)], \tag{13}$$

in which

$$\text{Var}(f_1) = (1/N_1) \sum_{j=1}^{N} (f_{1j} - \bar{f}_1)^2 \tag{14}$$

is the conditional variance of the probability judgments for the target event A on those N_1 occasions when it in fact occurs; $\text{Var}(f_0)$ has a similar definition and interpretation for the remaining N_0 occasions when the target event does not occur. $\text{Var}(f_1)$ and $\text{Var}(f_0)$ measure variability in the judge's assessments which is unrelated to whether or not the target event happens. So, from the perspective of anticipating that event, this variability is noise, or "scatter." $\text{Scat}(f)$, the weighted mean of $\text{Var}(f_1)$ and $\text{Var}(f_0)$, is thus an index of the overall scatter contained in the judge's *probability statements*.

The remaining term in the covariance decomposition of $\overline{\text{PS}}$ is $\text{MinVar}(f)$, which can be shown to be

$$\text{MinVar}(f) = \text{Var}(f) - \text{Scat}(f), \tag{15}$$

where $\text{Var}(f)$ is the variance of the entire collection of probability judgments for the target event. It can also be shown that

$$\text{MinVar}(f) = [\text{Slope}]^2[\text{Var}(d)], \tag{16}$$

which contains the elements of the covariance of judgments and outcome indexes. Accordingly, Eq. (15), which can be rearranged as

$$\text{Var}(f) = \text{MinVar}(f) + \text{Scat}(f), \tag{17}$$

can be seen as similar to a partition of variance in the analysis of variance, with $\text{MinVar}(f)$ analogous to effect variance and $\text{Scat}(f)$ corresponding to error variance. Since $\text{Var}(f)$ contributes to $\overline{\text{PS}}$ positively, ideally one would want to minimize it. But this would eliminate the slope, too. Thus, conditional upon the attainment of a given slope, $\text{MinVar}(f)$ reflects the amount of judgment variability that must be tolerated. That is why it is called the "*conditional minimum judgment variance.*"

THE MULTIPLE-EVENT COVARIANCE DECOMPOSITION

Now consider the multiple-event case. Let $A_1, A_2, \ldots, A_K$ constitute a K-event sample space partition, with $K \geq 2$. Then in the natural fashion we can define outcome indexes d_k for each event according to Eq. (1), $k = 1, \ldots, K$. Similarly, the respective probability judgments for $A_1, A_2, \ldots, A_K$ can be denoted by f_k, $k = 1, \ldots, K$. The outcome indexes and judgments can be represented compactly by vectors $\mathbf{d} = (d_1, d_2, \ldots, d_K)$ and $\mathbf{f} = (f_1, f_2, \ldots, f_K)$, respectively. So, in the labor expert example, $\mathbf{f} = (.60, .15, .15, .10)$. If the union rejects management's offer and strikes (event A_3), then $\mathbf{d} = (0, 0, 1, 0)$.

The *multiple-event probability score* (PSM) can be described as (cf. Murphy, 1972b)

$$\begin{aligned} \mathrm{PSM}(\mathbf{f},\mathbf{d}) &= (\mathbf{f} - \mathbf{d})(\mathbf{f} - \mathbf{d})' \\ &= \sum_{k=1}^{K} (f_k - d_k)^2 \\ &= \sum_{k=1}^{K} \mathrm{PS}(f_k, d_k). \end{aligned} \tag{18}$$

It is straightforward to show that $0 \leq \mathrm{PSM} \leq 2$. If i, with $i = 1, \ldots, N$, is used to index multiple-event judgments $\mathbf{f}_i$ and outcome indexes $\mathbf{d}_i$ over N different occasions, then the mean of PSM can be defined in the expected way:

$$\begin{aligned} \overline{\mathrm{PSM}}(\mathbf{f},\mathbf{d}) &= (1/N) \sum_{i=1}^{N} \mathrm{PSM}(\mathbf{f}_i, \mathbf{d}_i) \\ &= \sum_{k=1}^{K} \overline{\mathrm{PS}}(f_k, d_k) \\ &= \sum_{k=1}^{K} \overline{\mathrm{PS}}_k, \end{aligned} \tag{19}$$

where, to simplify the notation, $\overline{\mathrm{PS}}_k$ represents the mean probability score for the kth event in the partition. Thus, the sum of the mean probability scores for the individual events constitutes an overall measure of accuracy for the multiple-event judgments.[2]

A combination of Eqs. (4) and (19) yields the $\overline{\mathrm{PSM}}$ *covariance decomposition:*

$$\begin{aligned} \overline{\mathrm{PSM}}(\mathbf{f},\mathbf{d}) = &\sum_{k=1}^{K} \mathrm{Var}(d_k) + \sum_{k=1}^{K} \mathrm{MinVar}(f_k) \\ &+ \sum_{k=1}^{K} \mathrm{Scat}(f_k) + \sum_{k=1}^{K} \mathrm{Bias}_k^2 \\ &- 2 \sum_{k=1}^{K} [\mathrm{Slope}_k][\mathrm{Var}(d_k)], \end{aligned} \tag{20}$$

[2] The reader might note that the single-event situation is equivalent to the multiple-event situation in which the partition of the sample space consists of two events ($K = 2$), the target event and its complement. In that case, PS = (1/2)PSM.

in which the subscripts on the summands correspond to the respective individual events in the sample space partition. Each sum on the right-hand side of Eq. (20) has a meaning which generalizes the interpretation of the individual-event summands discussed previously. For example, normally $\sum_{k=1}^{K} \mathrm{Var}(d_k)$ is a part of the overall accuracy measure which is outside the judge's influence. And the sum of squared biases indexes the general extent to which the judge's assessments are miscalibrated.

As indicated above, Murphy (1972b) has derived other decompositions of what is called here $\overline{\mathrm{PSM}}$. Yates (1982) has shown what the relationship is between the single-event "new" Murphy (1973) decomposition of $\overline{\mathrm{PS}}$ and the covariance decomposition of the same. The relationships between the corresponding decompositions of $\overline{\mathrm{PSM}}$ follow directly. It is probably inappropriate to say that any probability score decomposition is in every respect "better" than the others. Each offers a different perspective on judgment accuracy that can inspire insights that are not encouraged by the alternatives. That is why any given accuracy study would do well to apply more than one analytical technique. The following illustrations serve to highlight the particular kinds of understanding prompted by the covariance decomposition.

ILLUSTRATION 1: MEDICAL DIAGNOSES

Habbema, Hilden, and Bjerregaard (1978) reported a collection of probability assessments made by a statistical model in a medical situation. The task was to diagnose the cause of patients' complaints about stomach ailments. In each instance, probability "judgments" were offered for three diagnoses: A_1 = "Nonspecific abdominal pain," A_2 = "Acute appendicitis," and A_3 = "Other diseases." Fifty cases were considered. Because the cases were so few in number, it should be recognized that the analyses described here are for demonstration purposes only.

Table 1 shows the $\overline{\mathrm{PS}}$ and $\overline{\mathrm{PSM}}$ covariance decomposition statistics for the alternative diseases. It is useful to have standards of comparison for values of $\overline{\mathrm{PS}}$ and $\overline{\mathrm{PSM}}$. One standard is the performance of the "uniform judge." The uniform judge reports that all events in the given partition are equally likely; i.e., $f_k = 1/K$, for $k = 1, \ldots, K$. For example, in the present illustration, the uniform judge would indicate that P' (Nonspecific abdominal pain) = P' (Acute appendicitis) = P' (Other diseases) = 1/3. The covariance decomposition of $\overline{\mathrm{PSM}}$ permits the direct conclusion that, for the uniform judge, $\overline{\mathrm{PSM}} = 1 - 1/K$. So, in the present diagnostic

TABLE 1
$\overline{PS}$ AND $\overline{PSM}$ COVARIANCE DECOMPOSITION TERMS FOR DISEASE PROBABILITY JUDGMENTS

Term	Sum	Event/Disease A_1 Nonspecific abdominal pain	A_2 Acute appendicitis	A_3 Other diseases
$\overline{PS}_k$	.2735[a]	.1131	.0333	.1271
Var (d_k)	.5704	.1924	.1344	.2436
MinVar (f_k)	.1500	.0281	.0700	.0519
Scat (f_k)	.1154	.0388	.0225	.0541
Bias$_k^2$	.0037	.0009	.0004	.0024
-2[Slope$_k$][Var(d_k)]	$-.5660$	$-.1471$	$-.1940$	$-.2249$

[a] $\overline{PSM}$ (f,d).

situation, the judge could assure $\overline{PSM} = 2/3$ by simply considering each of the possibilities to be equally likely. As indicated in Table 1, the assessments recorded by Habbema *et al.* surpassed the standard set by the uniform judge. This may seem like a minor achievement. But in a study of judgments concerning stock price activity, Staël von Holstein (1972) found that only 3 of 72 subjects outperformed the uniform judge.

The "base rate judge" offers a second, more exacting comparison standard. The base rate judge reports that the probability of each event is equal to its base rate; i.e., $f_k = \bar{d}_k$, $k = 1, \ldots, K$. The base rates for the respective alternatives were $\bar{d}_1 = 26\%$ for nonspecific abdominal pain, $\bar{d}_2 = 16\%$ for acute appendicitis, and $\bar{d}_3 = 58\%$ for other diseases. These would be the corresponding probability judgments for the base rate judge, too. Of course, one would have to be clairvoyant to know the exact value of $\bar{d}_k$ before the events actually occur. However, suppose a large data base is available, such that the relative frequencies of the events are very stable, near some value δ_k, for $k = 1, \ldots, K$. Then, for any reasonably sized sample of occasions, $\bar{d}_k$ should be close to δ_k also. That is, δ_k could be taken as a good estimate of $\bar{d}_k$.

An inspection of the $\overline{PS}$ decomposition makes it clear that, for the base rate judge, $\overline{PS}_k = \text{Var}(d_k)$. It then also follows that $\overline{PSM} = \sum_{k=1}^{K} \text{Var}(d_k)$. Table 1 shows that the predictions listed by Habbema *et al.* were superior to those of the base rate judge, both overall and for each disease individually.

Figure 1 shows an ensemble of "covariance graphs" for the events in the diagnosis sample space. Such displays afford visual appreciation of

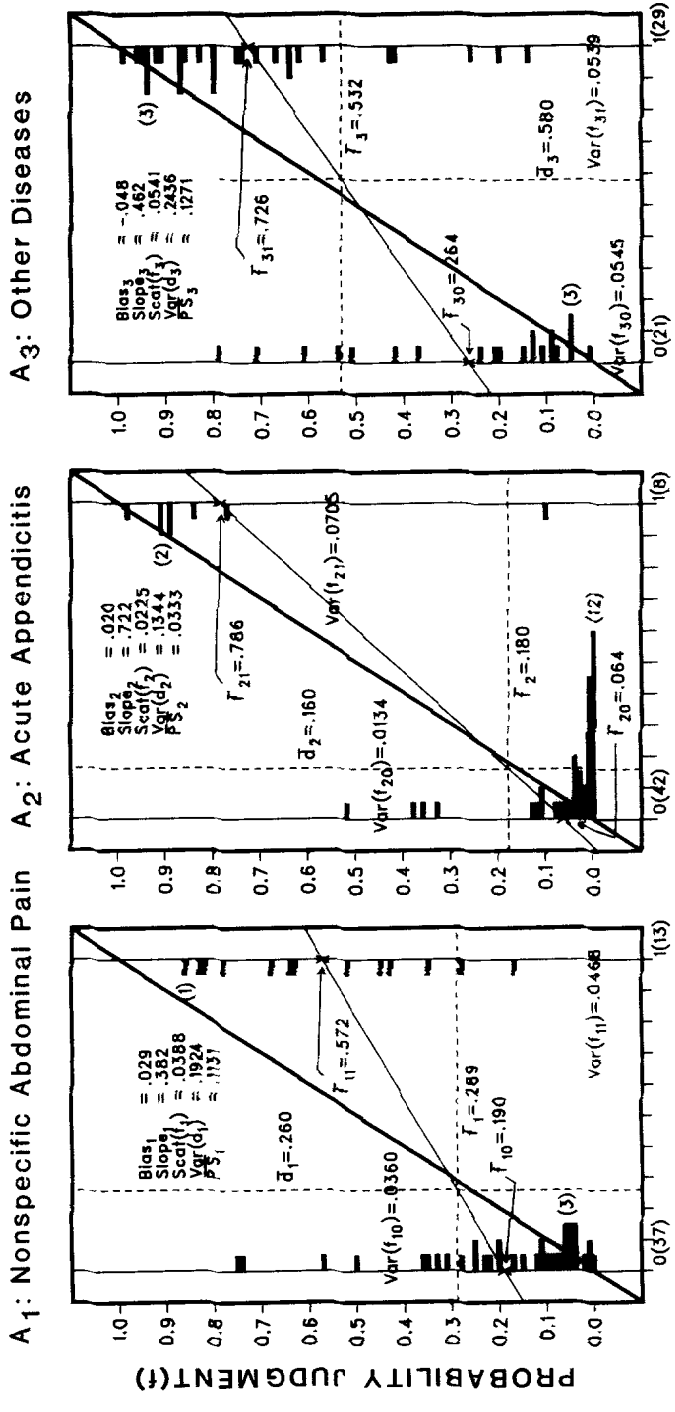

FIG. 1. Covariance graphs for disease probability assessments.

the judgment aspects indexed by the components of $\overline{PS}$ and $\overline{PSM}$ covariance decompositions. For convenience, relevant statistics are superimposed on the covariance graphs, too. Covariance graphs sometimes can be difficult to understand, especially when they are seen for the first time. Thus, the meanings of various features of such representations are discussed in some detail for the first covariance graph shown in Fig. 1, that for the diagnostic event "Nonspecific Abdominal Pain."

The abscissa in a covariance graph is defined by the outcome index d. Although the outcome index itself can assume only two values, 0 and 1, it proves to be convenient to mark off intermediate points between those extremes. It is customary to indicate in parentheses adjacent to the corresponding values of d how often the target event does and does not occur. Thus, it is shown in Fig. 1 that 37 of the 50 patients did not have nonspecific abdominal pain ($d_1 = 0$), while the remaining 13 did ($d_1 = 1$). The ordinate of the covariance graph is identified with the probability judgment f for the target event.

A covariance graph contains two histograms, one for probability judgments for the target event when it actually occurs ($d = 1$), the other for when it does not ($d = 0$). The scale of the histograms is indicated by labeling the longest bar in each with the number of cases symbolized, e.g., (3) for the histogram representing the 37 cases in which patients did not suffer nonspecific abdominal pain. Ideal judgment performance would be depicted by a covariance graph containing two degenerate histograms. The histogram on the right, above $d = 1$, would consist of a single bar at $f = 1$; that on the left, above $d = 0$, would have a solitary bar at $f = 0$. Roughly, accuracy is good to the extent that a covariance graph approaches this configuration. The various elements of the covariance decomposition, which are visually characterized by the graphical features distinguished below, are specific ways that real judgments fall short of the ideal.

Every covariance graph contains a horizontal dotted line through the overall mean probability judgment $\bar{f}_k$. For instance, Fig. 1 shows that $\bar{f}_1 = .289$; i.e., the average judgment for nonspecific abdominal pain was 28.9%, over all 50 patients. The covariance graph also includes a vertical dotted line through the base rate $\bar{d}_k$; e.g., $\bar{d}_1 = 26.0\%$ for nonspecific abdominal pain. The distance of the intersection of those lines from the 1:1 diagonal is the absolute value of Bias_k. For nonspecific abdominal pain judgments, the bias was thus $\text{Bias}_1 = +2.9\%$; the judgments were generally somewhat too high. Recall that $\bar{f}_{1k}$ and $\bar{f}_{0k}$ are, respectively, the mean probability judgments for the target event when it does and does not actually happen; e.g., $\bar{f}_{11} = 57.2\%$ and $\bar{f}_{10} = 19.0\%$ for nonspecific abdominal pain. The line passing through $(0, \bar{f}_{0k})$ and $(1, \bar{f}_{1k})$ in a covari-

ance graph is the regression line for judgments regressed on outcome indexes. The slope of that line is $\text{Slope}_k = \bar{f}_{1k} - \bar{f}_{0k}$. When accuracy is perfect, the slope is 1. For nonspecific abdominal pain, we see that $\text{Slope}_1 = .572 - .190 = .382$, a value far from perfection. The variability in the histograms on either side of a covariance graph is indexed by the conditional variances $\text{Var}(f_{1k})$ and $\text{Var}(f_{0k})$; recall that $\text{Scat}(f_k)$ is the weighted mean of these statistics. For nonspecific abdominal pain, these statistics are, respectively, $\text{Var}(f_{11}) = .0468$, $\text{Var}(f_{01}) = .0360$, and $\text{Scat}(f_1) = .0388$.

As indicated previously, bias, slope, and scatter are accuracy dimensions that are under a judge's control. Note that, as is apparent from the pertinent statistics and the covariance graphs, these aspects of performance were best for the diagnosis of appendicitis. Assuming that this is more than an unreliable statistical aberration, two plausible explanations suggest themselves. The first is that, because appendicitis is such a serious, life-threatening condition, physicians go out of their way to learn to diagnose it as well as possible—in every respect. The second hypothesis is that appendicitis is judged more accurately than "nonspecific abdominal pain" and "other diseases" because it is more precisely defined. Only more pointed empirical study with large numbers of cases can settle the question.

Another issue raised by the present illustration concerns the direction of the biases observed for the three diagnoses. Christensen-Szalanski and Bushyhead (1981) found very strong positive biases in physicians' judgments for the diagnosis of pneumonia. Centor, Dalton, and Yates (1984) recorded similar large biases in physicians' judgments for positive streptococcus test results for patients complaining of sore throat. An intuitively appealing hypothesis that in principle could account for these results is motivational: Since the physician does not want to "miss" diagnosing a serious condition in a patient, an inordinately high probability is offered for that condition. However, Christensen-Szalanski and Bushyhead (1981) found no support for this hypothesis when they tested it via questionnaires about perceived diagnostic error costs.

The complete PSM analysis of the present data suggests an additional hypothesis which should be pursued in further studies, even though it might not apply here, since the assessment came from a statistical model. Note that the observed biases were positive for the most well-defined diagnoses—acute appendicitis and nonspecific abdominal pain. Results reported by Koriat, Lichtenstein, and Fischhoff (1980) imply that common biases in probability judgments about general-knowledge questions are at least partly due to attention focusing. When faced with a two-alternative question, subjects appear to be overly influenced by ar-

guments favoring one of the alternatives, to the neglect of equally significant but less readily accessible arguments favoring the competing possibility. Perhaps the same principle applies in medical diagnosis. Suppose the physician is somehow led to consider a specific disease in a diagnostic situation. Then he or she is easily able to bring to mind reasons why that disease is the appropriate diagnosis. So a high probability is offered for that alternative, leaving little probability "left over" for catch-all categories, such as "other diseases": What is out of sight is out of mind.

ILLUSTRATION 2: GRADE PROJECTIONS

For counseling purposes, near the middle of each term some universities ask instructors to project the grade each of their students will earn. So it is of practical interest to know how well this task is accomplished. With this issue in mind, as well as the aim of demonstrating the usefulness of the multiple-event covariance decomposition in comparisons between judges, several university psychology instructors were requested to make multiple-event probability judgments for the final grades of students in their classes. The events were A_1 = "Will earn a grade of C+ or worse," A_2 = "Will earn a grade of B−, B, or B+," and A_3 = "Will earn a grade of A− or better." The judgments for each student were constrained to sum to 1.0. They were elicited in the approximate middle of the term, after the instructors had had the opportunity to examine one or more products of every student's work. Appropriate precautions were taken to protect students' identities.

Table 2 presents the values and covariance decomposition components of $\overline{PS}$ and $\overline{PSM}$ for the judgments made by Instructors 1 and 2. Instructor 1 was the individual who, among all those who participated, achieved the lowest, i.e, the best, $\overline{PSM}$ value; Instructor 2 obtained the highest. The sampling distributions of $\overline{PS}$ components have not yet been determined. So it is not possible to draw definitive conclusions about contrasts between the stable judgment tendencies of Instructors 1 and 2. Nevertheless, the present sample statistics can be used to discuss how comparisons could be made when samples are large enough to rely upon limit theorems to defend inferences. As in the medical example, the meaning of each statistic is enhanced by the ensembles of covariance graphs shown in Fig. 2.

First, note that, according to the scatter component of the $\overline{PSM}$ decomposition, Instructor 1 had less overall scatter in her judgments than did Instructor 2. This advantage is not immediately apparent in the dis-

TABLE 2
$\overline{PS}$ AND $\overline{PSM}$ COVARIANCE DECOMPOSITION TERMS FOR PROJECTED GRADE PROBABILITY JUDGMENTS

Term	Sum	Event/Grades A_1 (≤C+)	A_2 (B−, B, B+)	A_3 (≥A−)
		Instructor 1		
$\overline{PS}_k$	.3387[a,b]	.0449	.1693	.1245
Var (d_k)	.5949	.1071	.2498	.2380
MinVar (f_k)	.1712	.0290	.0510	.0913
Scat (f_k)	.1732	.0168	.0871	.0693
Bias_k^2	.0309	.0034	.0071	.0204
$-2[\text{Slope}_k][\text{Var}(d_k)]$	−.6309	−.1114	−.2257	−.2938
		Instructor 2		
$\overline{PS}_k$	.5858[a,b]	.1802	.2640	.1417
Var (d_k)	.6258	.2400	.2452	.1406
MinVar (f_k)	.0532	.0409	.0024	.0098
Scat (f_k)	.2265	.0972	.0640	.0653
Bias_k^2	.0017	.0002	.0011	.0004
$-2[\text{Slope}_k][\text{Var}(d_k)]$	−.3214	−.1981	−.0488	−.0744

[a] $\overline{PSM}$ (f,d).

[b] Due to rounding error, some horizontal and vertical sums are not exact.

persion of the histograms in the covariance graphs. Nor was that advantage uniform, in that Instructor 2's predictions were actually *less* scattered for two of the grade categories, as revealed by comparisons of the scatter components in the corresponding $\overline{PS}_k$ decompositions. One potential cause of scattering in probabilistic forecasts is that the judge responds to cues which are thought to be predictive, but which really are not. It is conceivable that Instructor 1 was less inclined to rely on misleading grade cues than was Instructor 2.

Observe that Instructor 1's judgment performance was not consistently superior to Instructor 2's in all components of $\overline{PSM}$. Although predictions of the former instructor were overall less scattered, her opinions were much more biased. This is revealed in the larger distances of the intersections of the $\bar{f}_k$ and $\bar{d}_k$ lines from the 1:1 diagonals in the respective covariance graphs, and is confirmed by the bias components of the decompositions. The graphs also highlight across-event differences in bias. Especially noticeable is how overly optimistic Instructor 1 was that her students would earn high rather than low grades. Perhaps the difference in bias was due to experience. Instructor 2 had taught her course much longer than had Instructor 1. So she may very well have developed a

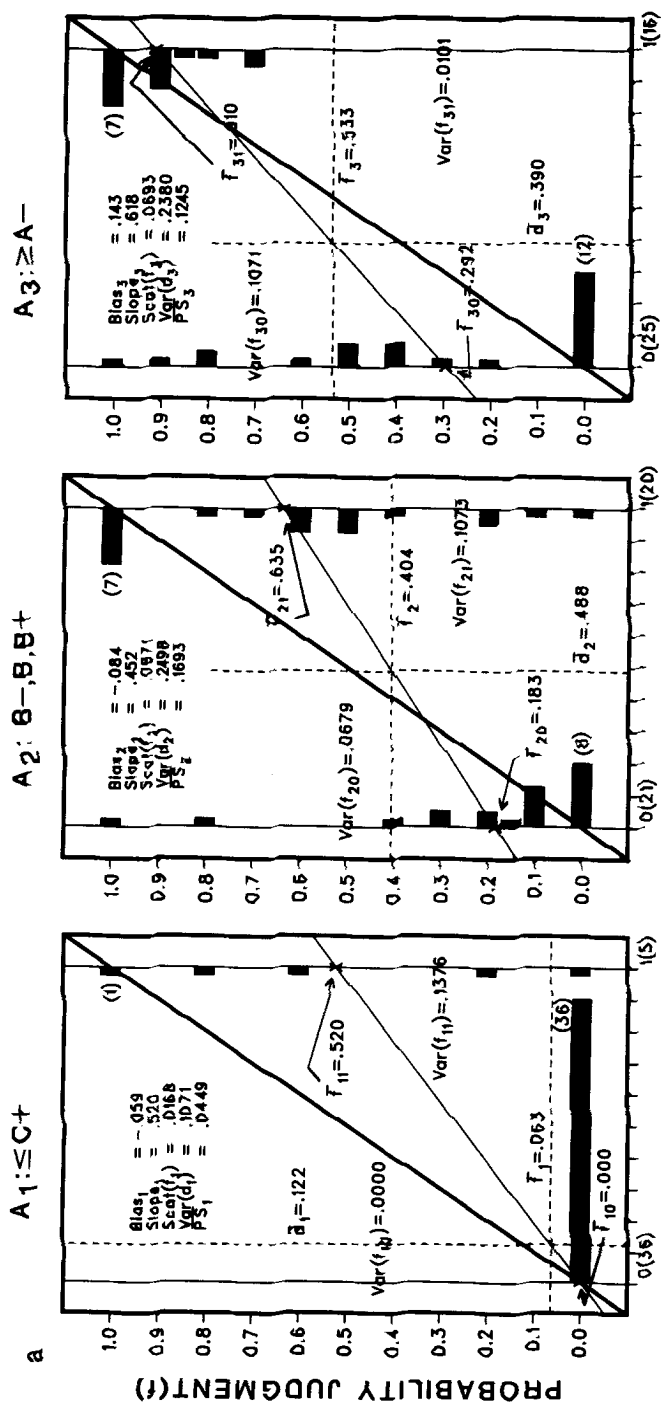

FIG. 2. Covariance graphs for grade projection probability judgments: (a) Instructor 1, (b) Instructor 2.

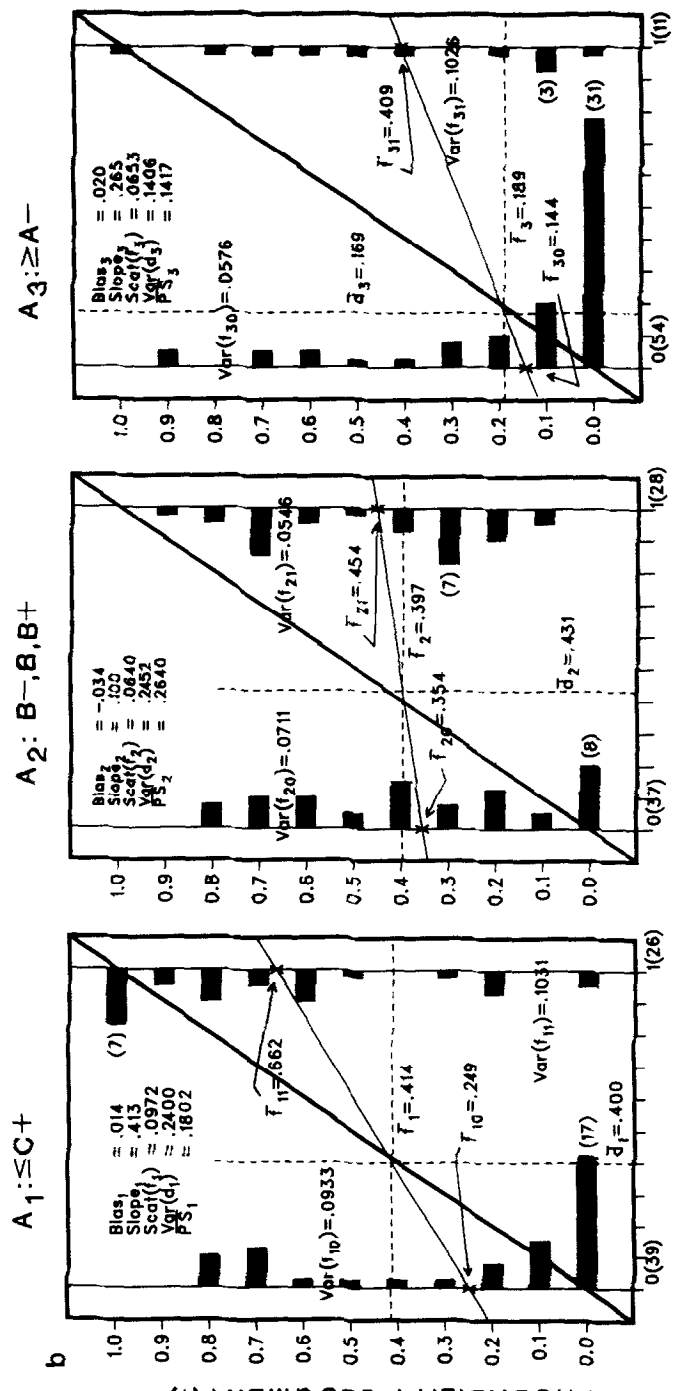

FIG. 2.—*Continued.*

better sense of how often students tend to earn A's, B's, and other grades.

The results show that the primary way Instructor 1 was able to achieve her overall superior accuracy was through judgment slopes. As indicated by the regression lines in the covariance graphs and by the slope components of the decompositions, for every event the slope for Instructor 1's judgments was greater than that for Instructor 2's predictions. A plausible explanation for this difference is detailed knowledge of each student's work. In general, a judge should be able to achieve a good slope only if he or she has access to predictive cues for every instance of an event and knows how to interpret those cues. Instructor 1 personally graded the assignments and tests for all the students in her class. However, Instructor 2 was aided by a teaching assistant who graded part of the course requirements. Thus, perhaps, her ability to make discriminative judgments was limited by the information available to her.

The outcome index variance term in the $\overline{PSM}$ covariance decomposition favored Instructor 1. This advantage was "undeserved," in that the outcome indexes were not determined by the instructors, but rather by the difficulty of the course material and the competence of the students. That is why a case can be made that a proper understanding of comparative judgment skills should not rely on gross accuracy measures like $\overline{PS}$ and $\overline{PSM}$. Instead, it should depend on comparisons of decomposition terms the judges are capable of affecting, e.g., bias, scatter, and slope. The separation of judge-controlled and judge-independent aspects of judgment accuracy is a major contribution of $\overline{PS}$ and $\overline{PSM}$ decompositions (cf. Murphy, 1973).

DISCUSSION

The illustrations presented above by no means exhaust the domains in which multiple-event decomposition methods should be useful. Another context is testing. For over 20 years, probabilistic responding to multiple-choice tests has been advocated by some observers (cf. Rippey, 1968; Shuford, Albert, & Massengill, 1966). In probabilistic or "confidence" testing, rather than categorically asserting that alternative (b) is the correct answer to a five-alternative, multiple-choice item, the testee would assign a probability judgment that *each* of those alternatives is the correct one. Decomposition analyses should be a valuable tool for probabilistic test item analyses. A closely related application is in the context of memory research, in which the subject must identify which of several items is old rather than new. Yet another application is in perception ex-

periments, where the subject is required to say, for instance, that a stimulus has occurred in one of several alternative locations.

It is important to recognize what procedures such as those described here can and cannot do. These techniques are useful in their ability to identify and quantify specific ways in which a person's judgmental accuracy is good or poor. It is then much easier to understand and correct shortcomings in that individual's judgment processes. An analogy is appropriate. Imagine a tennis player who is consistently losing. Tiring of this, she hires a coach to improve her game. After watching her play for a while, the coach announces, "You're not hitting the ball well. That'll be $30, please." Surely the player would refuse to pay for such vacuous, nonspecific "insight." Instead, she would demand a detailed analysis of exactly *how* she is hitting the ball improperly. Only then can she begin to make the required adjustments. Similarly, merely reporting an overall index of probability judgment accuracy can have only limited utility. Measures of specific, controllable accuracy dimensions, as provided by the covariance decomposition, offer a more promising avenue toward improvement.

When judgmental accuracy is deficient, decomposition statistics isolate the source of the problem in one or more narrow areas, e.g., bias or slope. This is certainly helpful. But an investigator might want even more pointed explanations of the difficulties. If so, he or she must go beyond what the statistics per se can provide. Suppose, for example, that a person's judgments contain an inordinate amount of scatter. There are at least two major reasons such a state of affairs might exist. The first is that the person executes his or her judgment policy in an inconsistent manner. The other is that, although the judgment policy is applied reliably, that policy relies on cues that are themselves only weakly related to the target event. Thus, as in the above medical and educational illustrations, while decomposition analyses can significantly reduce the number of plausible explanatory hypotheses, additional study is often necessary to arrive at definitive conclusions.

Although the present results are useful, there are several problems concerning multiple-event decomposition techniques which need to be addressed in further work. One of those problems concerns statistical inference. No one yet understands the sampling distributions of decomposition statistics under various conditions of general interest. So, until the required studies are performed, parametric inferences about the populations of those statistics must be very conservative, depending upon large samples. Or those inferences must come from less powerful nonparametric methods.

The illustrations in the present article involved probability judgments

about discrete events. There are many situations in which the objects of interest are quantities, e.g., commodity prices, sales totals, task completion times. In such contexts, the subject reports a judgment for the entire distribution for the quantity, perhaps via the fractile method or some other procedure (cf. Seaver, von Winterfeldt, & Edwards, 1978; Spetzler & Staël von Holstein, 1975). Staël von Holstein (1972) has illustrated how the accuracy of such distribution judgments can be indexed with a measure that is equivalent to $\overline{\text{PSM}}$. This is done by partitioning the continuum of possible values for the relevant quantity into successive intervals, e.g., $A_1 = (-\infty, x_1]$, $A_2 = (x_1, x_2]$, $A_3 = (x_2, x_3]$, etc. Ordinarily, the only aspect of distribution judgment accuracy which receives much attention is calibration (e.g., Alpert & Raiffa, 1982). As demonstrated, the methods described here provide measures of not only calibration, but other accuracy dimensions, too.

A recognized shortcoming of the probability score is its insensitivity to distance when applied to events that are naturally ranked relative to one another. As an example, suppose that $\mathbf{f}_1 = (.5, .2, .3)$ is a teacher's probabilistic forecast of the grade Student 1 will earn in a course, where A_1 = "Grade of C+ or worse," A_2 = "Grade of B−, B, or B+," and A_3 = "Grade of A− or better." Let $\mathbf{f}_2 = (.2, .5, .3)$ be the forecast for Student 2's grade. Suppose each student earns an A. Both forecasts would be assigned the same value of PSM, even though the mass of the probability distribution in the forecast for Student 2 was "closer" to the actual grade earned. There exist measures, such as the "ranked probability score" (Epstein, 1969; Murphy, 1971) and the "continuous ranked probability score" (Staël von Holstein, 1977), which are in fact responsive to distance. An important challenge for future efforts is to derive decompositions of measures like these which are comparable to the covariance decomposition of $\overline{\text{PSM}}$.

REFERENCES

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). New York: Cambridge Univ. Press.

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

Centor, R. M., Dalton, H. P., & Yates, J. F. (1984, November). *Are physicians' probability estimates better or worse than regression model estimates?* Paper presented at the Sixth Annual Meeting of the Society for Medical Decision Making, Bethesda, MD.

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

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8, 985–987.

Habbema, J. D. F., Hilden, J., & Bjerregaard, B. (1978). The measurement of performance

in probabilistic diagnosis: I. The problem, descriptive tools, and measures based on classification matrices. *Methods of Information in Medicine*, 17, 217–226.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 107–118.

Murphy, A. H. (1971). A note on the ranked probability score. *Journal of Applied Meteorology*, 10, 155–156.

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

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

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

Rippey, R. (1968). Probabilistic testing. *Journal of Educational Measurement*, 5, 211–215.

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

Seaver, D. A., von Winterfeldt, D., & Edwards, W. (1978). Eliciting subjective probability distributions on continuous variables. *Organizational Behavior and Human Performance*, 21, 379–391.

Shuford, E. H., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, 31, 125–145.

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

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

Staël von Holstein, C.-A. S. (1977). The continuous ranked probability score in practice. In H. Jungermann & G. de Zeeuw (Eds.), *Decision making and change in human affairs* (pp. 263–273). Dordrecht: Reidel.

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

Yates, J. F. (1984). *Evaluating and analyzing probabilistic forecasts* (UMAP Instructional Module No. 572). Lexington, MA: Consortium for Mathematics and Its Applications.

Yates, J. F., & Curley, S. P. (1985). Conditional distribution analyses of probabilistic forecasts. *Journal of Forecasting*, 4, 61–73.

RECEIVED: July 21, 1986