A Bayesian prediction of multiple-look identification performance from one-look data:

The effect of unequal prior probabilities*

STUART M. KEELEY and MICHAEL E. DOHERTY
Bowling Green State University, Bowling Green, Ohio 43403

Bayes's theorem had accurately predicted multiple-look visual identification performance from one-look data in two previous studies when the prior probabilities of the visual stimuli were equal. In this study, prior probabilities were manipulated to determine whether these values were important in determining agreement between man and statistical man. The prior probabilities of the stimuli, four gap orientations in Landolt rings, were .7, .1, .1, and .1 for four Ss and .3, .3, .3, and .1 for three Ss. One-look data were used to predict two-look and four-look performance. In the 7-1-1-1 condition, the empirical percent correct increased as a function of repeated observations for the rare stimuli, but not for the high-probability stimuli. In the 3-3-3-1 condition, Ss improved over repeated observations for both types of stimuli. Predictions of multiple-look performance under unequal prior probability conditions in this study were less accurate than those under equal prior probability conditions in previous studies. Three possible determinants of this difference were discussed.

Previous reports (Doherty & Keeley, 1969; Keeley, Doherty, & Bachman, 1970) have proposed Bayes's theorem as a model to describe the combination of multiple observations of briefly presented visual stimuli. When Ss' multiple-look data have been predicted from their single-look data matrices, Bayesian predictions of entire patterns of multiple-look identification performance have been quite accurate.

In applying the Bayesian model, we have assumed that S makes an implicit response to each observation within a multiple-observation trial and that the probability of an implicit response given each of the specified stimuli, $p(e_i/H_j)$, could be estimated by relative frequencies in the one-look condition. These conditional probabilities provide the basic input data for the predictions. For the model as we are using it, to predict S's multiple-observation performance, not only must the $p(e_i/H_j)$ values accurately reflect S's evaluation of the perceptual evidence provided by each observation in the multiple-observation condition, but these conditional probabilities must also be combined independently.

In the studies cited, the prior probabilities, $p(H_j)$, have been equal for the several j. It may be that the prior probabilities are important determinants of decision-making behavior, especially with respect to the question of whether judges can combine observations optimally. Thus,

this study manipulated the prior probabilities of occurrence of the visual stimuli. In requiring Ss to take into account the unequal prior probabilities, this study thus provides a further test of Bayes's theorem as a general model of visual identification performance.

METHOD

Subjects

Six undergraduate students served as paid Ss. In addition, the first author was run as a S.

Apparatus

A Scientific Prototype Model GB tachistoscope, with a handswitch permitting S to initiate stimulus onset, was the basic apparatus. The handswitch initiated two Hunter timers. One was wired into the circuit to keep presenting stimuli at .65-sec intervals, and the second controlled the total duration of the first and, therefore, the number of observations permitted S.

A response panel with a 4 by 4 array of buttons was placed before the Ss. The columns represented the four stimuli; the rows represented the four confidence ratings, indicating whether S "felt sure," "thought so," "was guessing," or "saw nothing." Ss responded by pushing the response button corresponding to both the stimulus he thought had been presented and to his confidence rating. The response panel was wired to a corresponding panel under E's control. Prior to a trial, E set the correct button, such that when S pushed the button S received immediate feedback of the correct response in the form of

a light coming on in the column representing the correct stimulus. A light on E's panel indicated which response S had made.

Stimuli

A black line, subtending a visual angle of approximately .2 deg, on a white background served as a fixation point. The white background and fixation point were illuminated at all times. The stimulus forms were Landolt rings with the gaps right, left, up, or down (stimuli will be represented by uppercase R, L, U, or D and responses by lowercase r, l, u, or d). The diameter of the rings subtended .3 deg of visual angle. The rings were centered approximately .3 deg above the fixation point. A single square stimulus slide was used, and E simply varied the orientation of the slide in presenting the four different gap positions.

Procedure

Viewing was monocular with S's preferred eye. All Ss participated in from four to eight pretraining sessions of approximately 1 h per session. During these sessions, prior probabilities were equal, and only single-observation trials were presented. For each S an exposure duration was determined at which a stable hit rate (HR) of approximately .50 was obtained for two consecutive sessions. This exposure duration was then maintained for all experimental sessions. Following pretraining, Ss were run from 15 to 20 experimental sessions, each session being approximately 1 h in length. A breakdown of the equipment prevented all Ss from completing the planned 20 sessions. Each session consisted of 10 practice trials and 75 experimental trials. During experimental sessions the prior probabilities of the four stimuli were .7, .1, .1, and .1 for four Ss and .3, .3, .3, and .1 for three Ss. The stimulus corresponding to the .7 prior probability was different for each S in the 7-1-1-1 condition, as was the case with the stimulus corresponding to the .1 prior probability in the 3-3-3-1 condition. Prior to the first experimental session, S was told the prior probability values of the stimuli that were to be presented to him. In addition, S was instructed to respond on the basis of what he thought was the most likely stimulus underlying the evidence he had obtained on a given trial. Thus, Ss were given the problem-solving instructional set to respond with what they thought was "out there." The S always responded by pushing a response button, indicating whether the gap orientation was right, left, up, or down and his confidence rating.

*This investigation was supported by PHS Research Grant No. MH-17487-01 from the National Institute of Mental Health.

Table 1
Obtained and Predicted Percent Correct (HR and HR') for Each S to .1 Stimuli and the .7 Stimulus in the 7-1-1-1 Condition for the 1C, 1C2, and 1C4 Trials. Each .1 value is a mean of conditional probability values for the three .1 stimuli for a given S.

Order of Prediction Matrix			1C	1C2	1C4
.1 Stimuli					
1	4 x 4	HR	.10	.19	.38
		HR'		.34	.53
2	4 x 5	HR	.29	.33	.44
		HR'		.44	.56
3	4 x 8	HR	.18	.25	.39
		HR'		.35	.43
4	4 x 4	HR	.24	.38	.53
		HR'		.41	.63
Mean		HR	.18	.28	.44
		HR'		.38	.54
.7 Stimulus					
1	4 x 4	HR	.90	.93	.84
		HR'		.82	.67
2	4 x 5	HR	.77	.73	.80
		HR'		.66	.62
3	4 x 8	HR	.75	.72	.71
		HR'		.56	.65
4	4 x 4	HR	.80	.86	.87
		HR'		.64	.41
Mean		HR	.81	.81	.81
		HR'		.67	.59

All seven Ss were run under three kinds of trials, randomized within each experimental session: one opportunity to observe (1C), two opportunities to observe (1C2), and four opportunities to observe (1C4). The S never knew the number of observations prior to the experimental trial. The number was programmed by E prior to the trial by setting the Hunter timer controlling the total duration of a trial. Stimulus presentation in all conditions was initiated by S after a ready signal, which was the sound of the stimulus tray being inserted into the tachistoscope. The S always emitted a single response following the last observation of a trial.

Prediction Model

The prediction model was used in the manner described in previous studies (Doherty & Keeley, 1969; Keeley, Doherty, & Bachman, 1970). Each S's 1C conditional probability matrix was used to predict his 1C2 and 1C4 matrices. The responses made in the 1C condition were assumed to represent the implicit responses available to S on each observation of the multiple-observation conditions. The conditional probabilities of each implicit response given each of the stimuli, $p(e_i/H_j)$, derived from the frequencies in the 1C condition

continue to represent the relative diagnostic impact of the implicit response, as they did previously when prior ps were equal. However, they now reflect weighting of unequal prior probabilities by the Ss. As before, it was assumed that S would, at the conclusion of observations on a trial, be in a perceptual state, D, determined by some combination of the implicit responses and that his decision would be determined by the most probable state of the world, given that perceptual state, $p(H_j/D_i)$, as calculated by Bayes's theorem. Since the weightings of the unequal priors should already be reflected in the $p(e_i/H_j)$ values, the calculation of $p(H_j/D_i)$ values for determining the optimal response simplifies to

$$p(H_j/D_i) = \frac{p(D_i/H_1) + p(D_i/H_2) + p(D_i/H_3) + p(D_i/H_4)}{p(D_i/H_1) + p(D_i/H_2) + p(D_i/H_3) + p(D_i/H_4)} \quad (1)$$

(see Doherty & Keeley, 1969, for a more detailed presentation of our use of the model.)

Although Ss were instructed to use confidence ratings, response frequencies across categories were usually insufficient to provide stable differential estimates of $p(e_i/H_j)$ values, which represented meaningful differences in diagnosticity among rating categories to a given identification response. This was especially true for the stimuli occurring on only .1 of the trials. Predictions were made for all Ss using the 4 (stimuli) by 4 (responses) prediction matrix. For three Ss, predictions were based on prediction matrices including confidence ratings, as well as the 4 by 4 matrix. For one S in the 3-3-3-1 condition, a 4 by 7 prediction matrix was used; and for two Ss in the 7-1-1-1 condition, 4 by 5 and 4 by 8 prediction matrices were used. The number of responses in the prediction matrix was determined by the Es on the basis of the number of rating categories S appeared to be using discriminatively. For example, in the 7-1-1-1 condition, one S used two confidence categories which differed in diagnostic impact for the response corresponding to the .7 stimulus but only one category for the other three responses. In the case of this S, a 4 by 5 prediction matrix was used for predictions. The 4 by 7 matrix in the 3-3-3-1 condition consisted of two confidence categories within the .3 responses and one category for the .1 response. The 4 by 8 matrix in the 7-1-1-1 condition included two confidence categories within each response.

Table 2
Obtained and Predicted HRs for Each S to .3 Stimuli and the .1 Stimulus in the 3-3-3-1 Conditions for the 1C, 1C2, and 1C4 Trials. Each .3 value is a mean of conditional probability values for the three .3 stimuli for a given S.

Order of Prediction Matrix					
S			1C	1C2	1C4
.3 Stimuli					
5	4 x 4	HR	.55	.68	.71
		HR'		.60	.75
6	4 x 4	HR	.50	.55	.62
		HR'		.48	.60
7	4 x 7	HR	.52	.65	.69
		HR'		.55	.65
Mean		HR	.52	.63	.67
		HR'		.54	.67
.1 Stimulus					
5	4 x 4	HR	.57	.79	.76
		HR'		.60	.68
6	4 x 4	HR	.29	.36	.23
		HR'		.49	.51
7	4 x 7	HR	.16	.32	.64
		HR'		.36	.48
Mean		HR	.34	.49	.54
		HR'		.48	.56

RESULTS

7-1-1-1 Condition

Table 1 presents for each S the mean obtained and predicted hit rates (HRs) for the 7-1-1-1 condition stimuli on the 1C, 1C2, and 1C4 trials. The .1 stimulus values are based on the average HR for the three .1 stimuli presented to a given S.

Empirically, all Ss showed a marked increase in HR as a function of increasing observations for the .1 stimuli. However, HRs remained relatively constant for the .7 stimulus as the number of observations increased.

Predicted values in Table 1 are based on the highest order 1C prediction matrix determined for a given S in the manner described above. For all Ss the prediction model overpredicted the mean HR to the .1 stimuli and underpredicted the HR to the .7 stimulus for both the 1C2 and 1C4 trials. The model predicts a gradual decrease in HR with increasing observations to the .7 stimulus and a marked increase to the .1 stimuli as a function of increased observations.

3-3-3-1 Condition

Table 2 presents the mean obtained and predicted HR for each S to the .3 stimuli and the obtained and predicted HR to the .1 stimulus for the 1C2 and 1C4 trials.

The empirical HR to the .3 stimuli increases monotonically as a function of increased observations for all three Ss. Results are less consistent to the .1 stimulus. For two Ss the HR increases

Table 3
Frequency and Conditional Probability of Each Response Given Each Stimulus, $p(e_i/H_j)$
for the 1C Condition for One S Each in the 7-1-1-1 and 3-3-3-1 Conditions

	Stimuli							
	S 2				S 5			
	R	L	U	D	R	L	U	D
	Prior Ps							
	.7	.1	.1	.1	.3	.1	.3	.3
r1	.48	.00	.00	.20	.00	.00	.00	.00
p(r1/H)	.16	.00	.00	.03	.00	.00	.00	.00
r2 + r3	.188	.15	.19	.30	.83	.14	.27	.43
p(r2 + r3/H)	.61	.42	.51	.51	.57	.26	.14	.30
l	.31	.15	.03	.10	.29	.30	.14	.10
p(l/H)	.10	.42	.08	.17	.20	.57	.07	.07
u	.29	.05	.13	.11	.18	.06	.130	.32
p(u/H)	.09	.14	.35	.19	.12	.11	.68	.22
d	.12	.01	.02	.06	.16	.03	.19	.59
p(d/H)	.04	.03	.05	.10	.11	.06	.10	.41

and then decreases as observations increase. The third S shows marked increases from one to two and from two to four observations.

As shown in Table 2, the prediction model underpredicts the HR to the .3 stimuli in the 1C2 trials but predicts quite accurately the 1C4 HRs. The predicted HRs to the .1 stimuli deviate more markedly from the empirical HRs than do the predictions to the .3 stimuli. The discrepancies are not systematic, and the magnitude of the discrepancies is similar in both 1C2 and 1C4 conditions.

DISCUSSION

Predictions of multiple-look visual identification performance from one-look data under unequal prior probability conditions were markedly less accurate than when the same stimuli were presented under equal prior probability conditions. The more priors deviated from equality, the less accurate were the predictions. Thus, the presence of unequal prior probabilities appears to represent one condition in the fixed observation procedure applied to visual identification in which man and statistical man do not agree. The salient question is, why?

An examination of the assumptions of the Bayesian model as we are using it provides several possible explanations for the finding of accurate predictions with equal prior probabilities and less-accurate predictions with unequal prior probabilities.

First, the model assumes that the $p(e_i/H_j)$ values obtained in the 1C condition for each response category estimate that response category's diagnostic impact for each stimulus state in the multiple-observation condition. The stability of these estimates is open to question in this study. With unequal prior

probabilities, the structure of the 1C matrices was quite different from that with equal prior probabilities. Table 3 presents representative examples of one-look data for one S each from the 7-1-1-1 and 3-3-3-1 conditions. In the 7-1-1-1 condition, most 1C responses corresponded to the .7 stimulus as would be expected, and only in this response category did Ss make many responses to more than one rating category. For S 2 in Table 3, the most common response was r. This S used three rating categories in responding r: right "1" (r1), right "2" (r2), and right "3" (r3). These right responses were collapsed into two categories in making the predictions to increase the stability of the $p(e_i/H_j)$ values. This collapsing of rating categories was necessary for all 7-1-1-1 Ss, and was due primarily to the low frequency of .1 stimulus presentations. As can be seen clearly from the table, small changes in frequency in the .1 stimulus cells might lead to marked changes in the measured diagnosticity of a response and, hence, marked changes in predictions. This was especially true for responses that showed little diagnostic impact. In the 7-1-1-1 condition, the resultant effect of the necessary collapsing of rating categories was the presence of one dominant response in the prediction matrix that was the most frequent and had the least measured diagnostic impact of all responses. In the case of S 2, this was the r2 response, since r2 was assigned proportionately very often to all four stimuli. The resultant effect of this 1C prediction matrix structure is a large proportion of perceptual states (D_i) containing the dominant implicit response, r2 in the case of S 2. For this S, when r2 occurs in combination with l, u, and d implicit responses, only when all implicit responses making up a D_i are r2 will the model predict a right

response, because all other response categories have substantially greater measured diagnostic impact. Thus, to the extent that the 1C prediction matrix has a very common response (due to the unequal priors in this case) but one which has relatively nondiagnostic rating categories, underprediction of the common responses and overprediction of rare responses will occur. This was the case for all 7-1-1-1 Ss. Furthermore, the decrease in HR predicted to the .7 stimulus as a function of increased observations for all four 7-1-1-1 Ss would be expected from a 1C matrix with the above-mentioned structure.

Related problems in the structure of the 1C matrix occur in the 3-3-3-1 condition (see Table 3), although not to the extent found in the 7-1-1-1 condition. Sampling errors would be expected to have had less effect because of the greater symmetry of the matrix. However, the major problem in this condition was that only one S used multiple rating categories with appreciable frequency in this condition, reducing the sensitivity of the prediction matrices.

A second assumption of the model is that Ss aggregate the evidence optimally. Studies of subjective probability revision indicate that Ss tend to revise conservatively, the magnitude of subjective revision being less than the data warrant (Edwards, 1968). A study by Peterson and Miller (1965) indicates that Ss are most conservative when the diagnostic impact of the evidence is large and revise their subjective probability estimates more than the amount specified by Bayes's theorem when evidence impact is slight. With the procedures used in the present study, it is difficult to determine what effect nonoptimal revision would have on a S's multiple-observation decision matrix. While Peterson and Miller utilized data with known diagnosticities, the present procedure requires inference of $p(e/H)$ values from the 1C decisions. If one assumes that the 1C matrices are direct measures of the subjective evidence impact, which assumption may be challenged, then the results obtained in the present study would be expected from the Peterson and Miller findings. If Ss overrevise their opinions when evidence that has little diagnostic value favors the dominant stimulus and underrevise when highly diagnostic evidence favors the nondominant, this would lead Ss to make more decisions to the dominant stimuli in the multiple-observation conditions than predicted.

A third assumption of the model is that observations are independent. With our procedures, perceptual independence has to be inferred

indirectly from the goodness of fit of the model (see Garner & Morton, 1969). Two previous studies (Doherty & Keeley, 1969; Keeley, Doherty, & Bachman, 1970) have predicted well, assuming independence. In a more recent study (Keeley & Doherty, in preparation), we have provided a very stringent, fairly direct test of the assumption, and the evidence strongly favors independence of observations. Furthermore, in a rather direct test of a specific alternative (Haber, 1967) to the independence assumption, the alternative has been rejected (Doherty & Keeley, in preparation). Thus, it is doubtful that inaccurate predictions in this study reflect a lack of perceptual independence.

The above procedures involve both the perceptual and decision processes of the human observer. The authors have been interested in determining under what conditions Ss combine information optimally and under what conditions they fail. This question places major emphasis on the decision

processes. Although the fixed observation methodology lends itself well to studying how Ss combine visual information, the problems encountered in this study suggest that, while this procedure is a theoretically appropriate one, it is a practically inefficient one for studying conditions which affect decision processes. Hours of pretraining and then a very great number of hours of S running are required to develop stable conditional probabilities, especially when the probability of one or more stimuli is low, as in the present study. In addition, many assumptions must be made about the IC matrix as an adequate indicator of conditional probability values. Also, the use of decisions as our sole dependent variable ignores other useful information, such as subjective probability revisions, which might shed light on the above-mentioned processes, although this other information could be collected in this situation. For these reasons, the authors feel the above question can be

answered more efficiently by other procedures.

REFERENCES

- DOHERTY, M. E., & KEELEY, S. M. A Bayesian prediction of four-look recognition performance from one-look data. *Perception & Psychophysics*, 1969, 5, 362-364.
- EDWARDS, W. Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment*. New York: Wiley, 1968.
- GARNER, W. R., & MORTON, J. Perceptual independence: Definitions, models, and experimental paradigms. *Psychological Bulletin*, 1969, 72, 233-259.
- HABER, R. N. Repetition as a determinant of perceptual recognition processes. In W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form*. Cambridge, Mass: M.I.T. Press, 1967.
- KEELEY, S. M., DOHERTY, M. E., & BACHMAN, S. M. A Bayesian prediction of four-look recognition performance from one-look data: II. *Perception & Psychophysics*, 1970, 7, 218-220.
- PETERSON, C. R., & MILLER, A. J. Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, 1965, 70, 117-121.

(Accepted for publication February 15, 1971.)