# A BAYESIAN METHOD FOR EVALUATING TRAINEE PROFICIENCY

KENNETH I. EPSTEIN and
FREDERICK STEINHEISER, JR.
Army Research Institute for
the Behavioral and Social Sciences

## Introduction

No instructional system is complete without a strong testing component. We hope that our instruction has been well enough designed so that any student who begins an instructional program will be able to achieve all of the objectives that the program was designed to teach. However, some students may require remedial or other supplementary instruction to master all of the objectives, even though the program was carefully developed. Furthermore, during the development of the instruction, test data from prospective students are required to first revise and later validate the instruction. In order to support the instructional development activities and to make decisions about the abilities of students who have completed instruction, a powerful testing program is necessary.

The final desired output of a test for a given examinee is information which allows us to pinpoint his ability to do whatever is required by an objective. That is, we observe a test score and must then infer the ability of the examinee. This paper outlines a "Bayesian" method for drawing such inferences. In addition, the adequacy of the method as a function of the number of test items administered and the effects of the tester's beliefs about the examinee population on the inferences drawn are discussed and illustrated.

Using the Bayesian method we were also able to hypothesize varying numbers of ability groups so that the classification of examinees into these ability groups is most useful to the overall instructional system. For example, the simplest case is to classify examinees into two groups, the first group containing those who have mastered the objective, and the second containing those who have not. Alternatively, one could hypothesize three groups, consisting of masters, nonmasters, and an intermediate group containing people whose skills are almost satisfactory and who could be brought up to the mastery level with relatively little additional instruction. The Bayesian model presented in this paper explores up to three levels of mastery, although this number could easily be expanded. The model also explores the effects on decision making (correctly classifying masters and nonmasters) if more than two ability levels have been hypothesized, but are then collapsed to form just two groups, of masters and nonmasters.

## Training to Mastery

Ideally, the educational decision-maker wants to know if a person (student, trainee) can do a job at some pre-specified level of acceptability. If a student's score on a test is above the minimal passing standard, he may be classified as a master; if his score is below the minimal passing score, he would be termed a nonmaster. But since data always have some error variability, misclassifications are likely to occur.

| | | True Competency State | |
|---|---|---|---|
| | | Master | Nonmaster |
| Classification Based on Test Score | Master | True positive | False positive |
| | Nonmaster | False negative | True negative |

Ideally, the probability of a True positive should be much greater than that for a False positive, and the probability for a True negative should be much greater than that for a False negative.

In order to evaluate how well our testing program achieves this goal, we want to be able to infer as accurately as possible the conditional probability of the mastery (or nonmastery) state, given the test score data, $p(M1|T)$, $p(M2|T)$. Our first problem is: What amount of data is this probabilistic inference based upon? Suppose that the passing standard was 80% of the test items correct. If a student had 33 out of 40 items correct, he would pass, and be classified as a master. Now suppose that on another form of the test (or a test given over the same material by another instructor), another student gets 25 out of 30 test items correct. This student would also have met the 80% correct criterion, and be classified as a master. The model presented in this paper will show that the p(M1|T) varies systematically with the number of test items, along with the minimal % correct for passing.

We may also ask: How is the accuracy of inference about mastery affected by postulating more than two states (mastery and nonmastery); and, can the data from various states be combined without seriously affecting the final p(M1|T) inference? For example, suppose that there are intermediate states of partial mastery. The following decision model shows that p(M1|T) can be more validly estimated when the mastery states are processed independently, but that the educational decision maker will not sacrifice very much classification accuracy if indeed he does dichotomize multichotomous data. Recall that we suggested that defining an intermediate group which required minimal remediation might be useful for some instructional systems. The model shows that the probability of being in the mastery group when indeed the datum was a test score obtained by a master will be increased if the other data are processed independently. The concept of "independent processing" requires that all nonmastery groups maintain their integrity, rather than being aggregated into one generalized nonmastery group.

## Bayes' Theorem

The statistical model which we have applied for classifying students into mastery and nonmastery groups, given their test score, is based upon a form of Bayes' Theorem:

$$p(M1|T) = \frac{p(T|M1)p(M1)}{[p(T|M1)p(M1) + p(T|M2)p(M2)]}$$

Here we assume that the 2 states of nature (master and nonmaster) are mutually exclusive and collectively exhaustive, and that T is the test score which is observed. We also assume that the test is dichotomously scored. A correct response is denoted "1", an incorrect response is denoted "0" and the total test score is simply the number of correct responses. What we seek to find is the term on the left, the probability that a given student is a master, having been given his test score. In order to find it, we need to have an estimate of the prior probability of mastery (p(M1)) in the population of students from which this student was drawn. The prior probability of mastery can be thought of as the proportion of students in the examinee population we think are masters. For example, if our instruction were very good the prior probability of mastery would be high, and most of the students who completed the instruction should have mastered the objective. The actual number specified for the prior probability of mastery may be an informed guess based on experience or it may be based on the empirical results of tests given to previous classes of similar students.

We must also estimate the conditional probability of a certain test score given that the student who got that score was a master. For example, if only one item were administered, the conditional probability of a score of one correct given that the student was a master is simply the probability that a master responds correctly. We may estimate this conditional probability empirically based on previous student groups, or we may provide a best guess as to how well masters perform, or this conditional probability may reflect a minimal standard of achievement. We shall show how the p(M|T) will vary as a function of the prior expectations of the tester, number of test items, and conditional probabilities, p(T|M), after an example to illustrate the computations.

Suppose that a student chosen at random from a trainee population was given a criterion-referenced test, and that he passed the test. Given the results of the test, what is the probability that the student is indeed a master of that particular course of instruction? In order to calculate the probability, we obtain the following information from the educational expert who administered the CRT: The probability that a master would obtain a passing score = .90, (p(T|M1) = .90); the probability that a nonmaster would obtain a passing score = .05, (p(T|M2) = .05); and the prior probability of randomly selecting a master from this trainee population is equal to .70; that is, we believe that 70% of this and similar previous trainee populations may be assumed to be composed of masters. Substituting these values into the formula

$$p(M|T) = \frac{.9 \times .7}{.9 \times .7 + .05 \times .3}$$

which equals .977. Hence, before the test score was available, the probability that this student was a master was .70, but after a passing score was observed, the probability that this person is a master has increased to .977. (The probability of this student being a nonmaster, given the same passing score, p(M2|T), would be equal to 1-.977 or .023.)

In order to generalize the Bayesian approach to a wide variety of applications in evaluating training effectiveness, two additions must be made to the basic formula. These additions are the number of trials or items on the test (N), and the number of hypothesized mastery states (S). The derivation of the general Bayesian formula for this purpose was originally presented by Hershman (1971):

$$p(Mi|T)=\frac{\prod_{j=1}^{N} p(Mi|t_j)}{p(Mi)^{S-1}\sum_{i=1}^{S}\frac{\prod_{j=1}^{N} p(Mi|t_j)}{p(Mi)^{S-1}}}$$

In this formula, p(Mi|tj) equals the conditional probability of a person in the ith mastery state getting the jth test item correct; p(Mi) is the prior probability of the representation of the ith mastery state in the student population (the % of students who are estimated to be in the ith mastery state); and p(Mi|T) is the conditional probability of a particular student being classified as being in the ith mastery state given his total test score. A computational example showing how the formula is applied for three mastery states is given in Appendix I.

## Variables of Interest in the Present Simulation

In the typical situation for evaluating training proficiency, the tester has some control over the number of items or trials that he will include on a test. In a performance-based test each trial may be rather expensive (such as tank gunnery or field artillery, where each shell costs over $100), and so the tester will be obliged to use a minimum number of trials to meet his decision-making requirements. Consequently, we examined the effect on p(M|T) when N took on values of 5, 10, 20, and 40 trials.

The tester also has some control over the values he assigns to the prior probabilities of mastery p(Mi), and the p(t|Mi) conditional probabilities. Values for both sets of probabilities were systematically varied in the present simulation.

The number of mastery states is a variable which the trainer and/or tester may also set. In some measurements of trainee proficiency it may be most appropriate to dichotomize on an all-or-none basis, whereas other training evaluation contexts may suggest a "pass, give refresher training, recycle failures through complete training" trichotomy. More than three mastery states may of course by hypothesized, but the computations in the present and all other models of proficiency evaluation become extremely complex. (However, we are developing a computer program which will handle up to five states of mastery.)

The dependent variable of main interest is the percent of items answered correctly. The tester may decide that 70% is a passing score. But the 70% value is not an absolute standard, since it is dependent upon the number of test items, and the prior and conditional probability estimates. In the present simulation, three values of per cent correct observed scores were used: 60%, 70% and 80%.

## Changes in p(M|T) Assuming Two Mastery States

The fundamental purpose of the present study was to investigate how the probability of mastery classification changes as a function of the simultaneous manipulation of up to four parameters (independent variables). The scope of the study is not exhaustive, since only several values of each of the four variables were used. However, some general trends do seem to emerge as can be seen in the following figures.

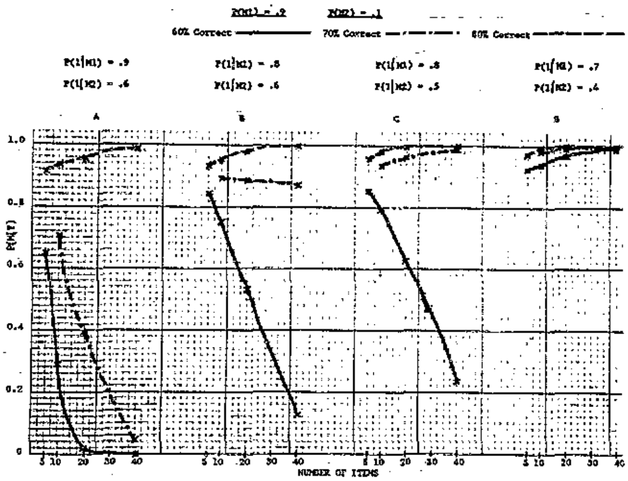


Figure 1.

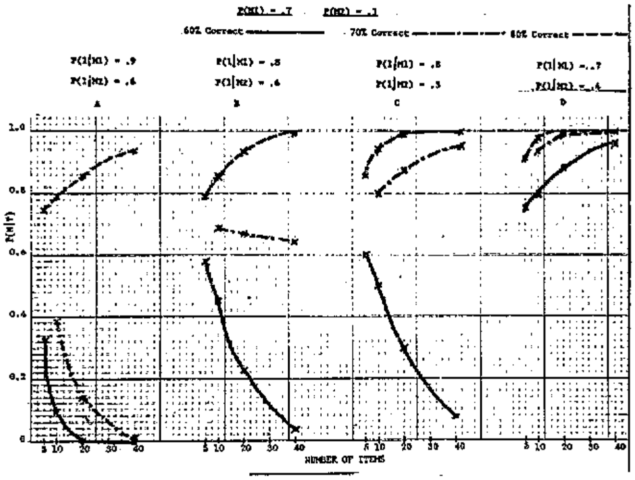


Figure 2.

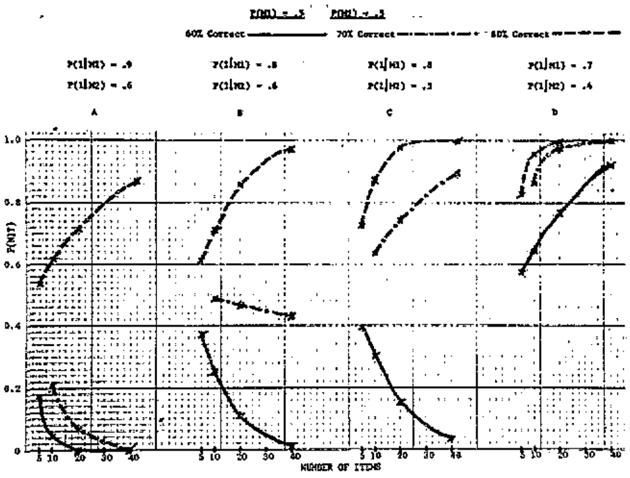


Figure 3.

Figures 1, 2, and 3 show the results of applying the model to a situation in which only two mastery groups (mastery and nonmastery) have been hypothesized. The data points represent the probability that a trainee is a master, given (conditional upon) his total test score, P(M|T). The curvature of each line shows how the P(M|T) changes as a function of variations in the prior expectation of mastery, the % correct items observed, the conditional probabilities of both a master and a nonmaster responding correctly to an item, and the number of items comprising the test.

Figure 1 represents a testing situation in which the training was of extremely high quality, since the proportion of masters in the trainee population was assumed to equal 0.9. That is, p(M1) = 0.9. Figure 1A portrays the situation in which both masters and nonmasters have attained a rather high degree of proficiency, since the probability of a master responding correctly to any given item is 0.9, and the probability of a nonmaster responding correctly is 0.6. If a person scored 80% on a five item test, the probability that he is a master is approximately .91. This probability drops to .65 if a 60% score on five items (3 out of 5 correct) were obtained. Note that when the test length is increased to 40 items, an 80% score (32 correct) produces a .99 probability of mastery. However, a score of 60% (24 correct) yields an essentially zero probability of mastery. The effect of the test length variable on classification accuracy is dramatic: if the p(M/T) had to be at least 0.5 for a person to be called a master, then scores of 60% on a <u>five</u>-item test would lead to <u>mastery</u> classification. But a 60% score on a <u>40</u>-item test would lead to <u>non</u>mastery classification.

Figure 1A also illustrates the effect of "prior beliefs" on p(M|T). Intuitively, one might suppose that the chances were much higher that a person who obtained a score of 60% (even from a 5-item test) came from a population whose probability of correctly answering an item was 0.6 than from a population whose probability of answering an item correctly was 0.9. However, the relative proportions of the two groups (expressed as prior belief in mastery and nonmastery, or p(M1) = .9 and p(M2) = .1, respectively) are such that the probability of a person being in the mastery state is approximately 0.65 for a score of 3 correct (60%) on a 5-item test. Only by increasing the number of test items can the strong prior bias in favor of the mastery decision be reversed. Figures 2A and 3A show what happens when prior beliefs are not so heavily biased in favor of mastery. In neither case is the probability of being in the mastery state above 0.5 for scores of less than 80%. But Figure 1A suggests that when prior beliefs heavily favor one group over the other, longer length tests should be used. Otherwise, the amount of data may not be sufficient to force a change in the originally held prior beliefs.

The effect of changing the prior beliefs concerning the proportion of masters and nonmasters in the examinee population while holding all other parameters constant can be seen by comparing corresponding graphs A, B, C, and D in Figures 1, 2, and 3. As the prior beliefs approach equiprobability (where $p(M1) = p(M2) = 0.5$), more items are required to maintain a given level of confidence that a person is either a master or nonmaster. The inability to postulate strong prior beliefs must be compensated for by increasing the test length in order to maintain a constant classification accuracy.

The effect of changing the probability of a correct response, $p(1|Mi)$, can be seen by comparing graphs A, B, C, and D for Figures 1, 2, and 3. For example, the only difference between Figure 1A and Figure 1B is that the $p(1|M1)$ changes from 0.9 to 0.8, all other parameters being held constant. (This change might reflect a lower level of required proficiency, and hence less training, for Graph B than for A. Or perhaps previous test results indicate that masters of the instruction respond to items with a probability of correct response equal to 0.8 rather than 0.9.) In any case, the effect of this small change in the $p(1|M1)$ on the $p(M|T)$ is readily apparent. For any test length or observed test score, the probability of being in the mastery state is greater in Graph B than in A. This shift is most obvious for the 70% observed correct curve. Notice that $p(M|T)$ on Graph A for an observed score of 70% (28 out of 40 correct) is approximately 0.04. However, the value for $p(M|T)$ in Graph B for 70% of a 40-item test correct is 0.87.

The main reason for this abrupt change from Graph A to B (in Figures 1, 2, and 3) is the lowered requirement for mastery, from 0.9 to 0.8. The probability that "0.9 persons" score only 70% correct on long tests is relatively low. But when masters are defined as those trainees who come from a population with a probability of responding correctly equal to 0.8, the probability of their scoring 70% on a long test is high. One of the most difficult jobs for an instructional designer is to describe the level of capability required of graduates and the level of capability actually achieved. Comparison of these graphs indicates the magnitude of the effect that these specifications can have on the classification of trainees.

Graphs C and D of Figures 1, 2, and 3 further illustrate the effect of variations in the probability of correct responses. The only difference between Graphs B and C is that the probability of a correct response from a nonmaster decreases from 0.6 to 0.5. The effect of this decrease in correct response probability from a nonmaster is to lower the likelihood of a nonmaster achieving a test score of at least 70%, which also increases the probability that a person achieving a high % score is in the mastery state. Finally, Graph D portrays an extreme case in which neither masters nor nonmasters are responding at particularly high levels. However, the level of performance for nonmasters is so low (0.4), that even for observed scores of 60% the probability of being in the mastery state exceeds 0.8 for all test lengths, except for 5 and 10 items in Figure 2, and 5, 10, and 20 items in Figure 3.

Further detailed analysis of these figures is not included in this paper. In comparing the twelve graphs against each other, note the magnitude of the changes in $p(M|T)$ when small changes have been made in the prior beliefs, in the correct response probabilities, and in the percent correct observed responses. The implication is that extreme care must be taken when specifying parameters in a Bayesian approach to testing and decision making. If the parameters are realistic, great savings in testing time and expense, and increased confidence in decision making are possible (Novick & Lewis, 1974). However, if the parameters are not realistic, there is a very real danger of misclassifying many examinees. The next section of this paper deals with an elaboration of the model to three mastery states, thus helping to quantify sources of classification error.

## Elaboration to Three Mastery States

Figures 4, 5, 6, and 7 represent cases for which three mastery states have been hypothesized. In figures 4 and 6 the probability of a correct response for a person assumed to be in mastery state M1 equals 0.8, for mastery state M2 this probability is 0.6, and for mastery state M3 it is 0.5. These values could correspond to the situation in which the nonmastery group was divided in half. That is, those persons whose probability of getting any given item correct is 0.5 (comprising mastery state M3) would need extensive retraining; whereas those whose probability is 0.6 (comprising mastery state M2) would merely need selective retraining. People in mastery state M1 have a probability of 0.8 for making a correct response, and may therefore be considered as "masters" who have successfully passed training.

For Figures 5 and 7 the corresponding probabilities of a correct response for people in mastery states M1, M2 and M3 are 0.9, 0.8, and 0.6, respectively. These probabilities might describe a situation in which the mastery group was dichotomized, perhaps in an attempt to identify those students who had achieved an exceptionally high level of proficiency, i.e., p(1|M1) = 0.9.

In Figures 4 and 5 the prior probability (or assumed proportion) of examinees in each mastery state are: p(M1) = 0.5, p(M2) = 0.3, and p(M3) = 0.2. In Figures 6 and 7 the corresponding prior probabilities are 0.25, 0.50, and 0.25, respectively. The prior values in Figures 4 and 5 display a bias towards higher levels of mastery (50% of the examinees are assumed to be type M1 masters), whereas the bias in Figures 6 and 7 is towards the intermediate level of mastery (50% of the examinees are assumed to be type M2 masters).

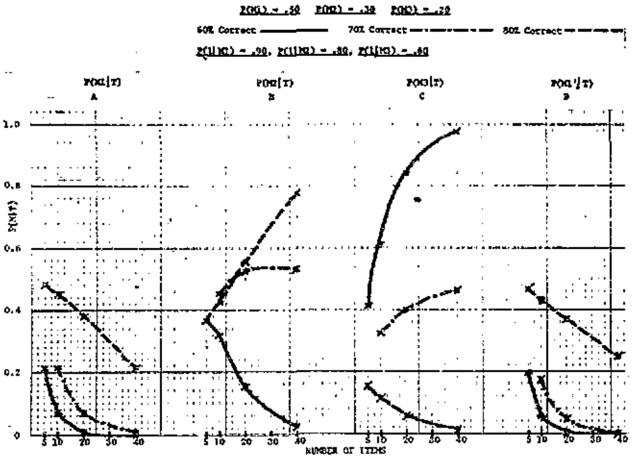


Figure 5.

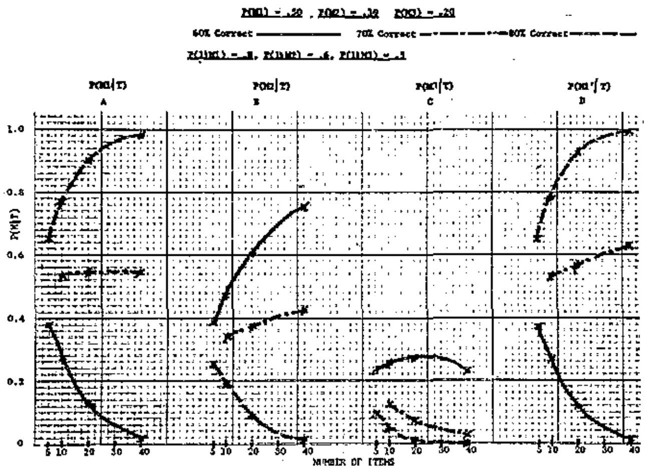


Figure 4.

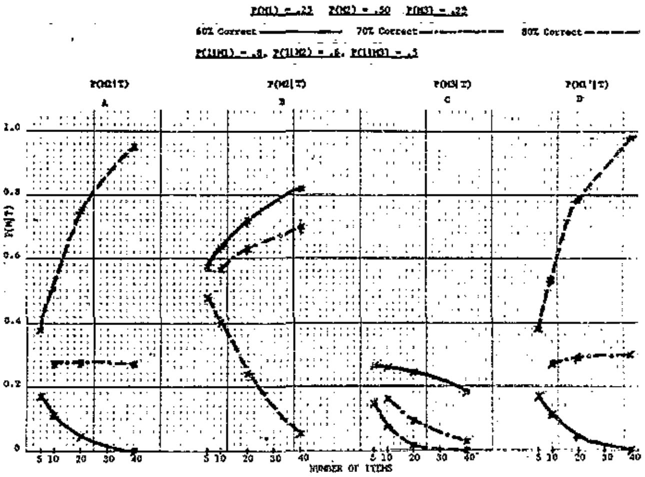


Figure 6.

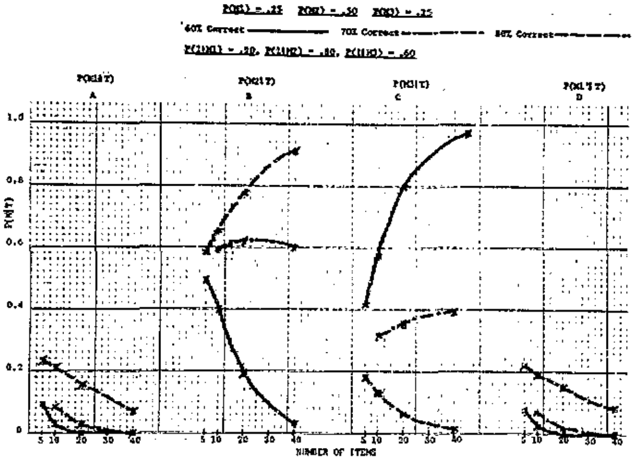


Figure 7.

A detailed analysis of Figures 4 and 5 will provide the basis for an interpretation of Figures 6 and 7, which is an exercise left to the reader. The three graphs labeled A, B, and C represent the probability that an individual is in mastery state M1, M2, and M3, respectively. Graph D represents the probability that a person is in mastery state M1 after mastery states M2 and M3 have been combined into one composite state.

Graph A shows the probability that an individual is in mastery state M1 given observed scores of 60%, 70%, and 80% correct on 5-, 10-, 20-, and 40-item tests. Thus, for an observed score of 4 out of 5 correct, the probability that this person is in mastery state M1 is about 0.65. But if this same person got a score of 32 out of 40 (still 80% correct), the probability that he is an M1 master jumps to 0.98. These results are similar to those obtained when two mastery groups were hypothesized, and again illustrate the effect of increasing test length on the level of confidence in the mastery classification p(M|T).

The probability of being in mastery state M2 given observed scores is plotted in Graph B. If a person got 4 out of 5 correct, the probability of being in state M2 is about 0.25. However, if he got 32 out of 40 correct (still 80% correct), this probability plummets to 0.02. Finally, using these same test score values, Graph C shows that the probability of being a type M3 master is 0.10 for 4 out of 5 correct, and nearly zero for 32 out of 40 correct. This result makes intuitive sense, because there is only 20% of type M3 (non)masters in the examinee population, and the probability of their getting any item correct is only 0.50, which is a long way from 80% observed correct.

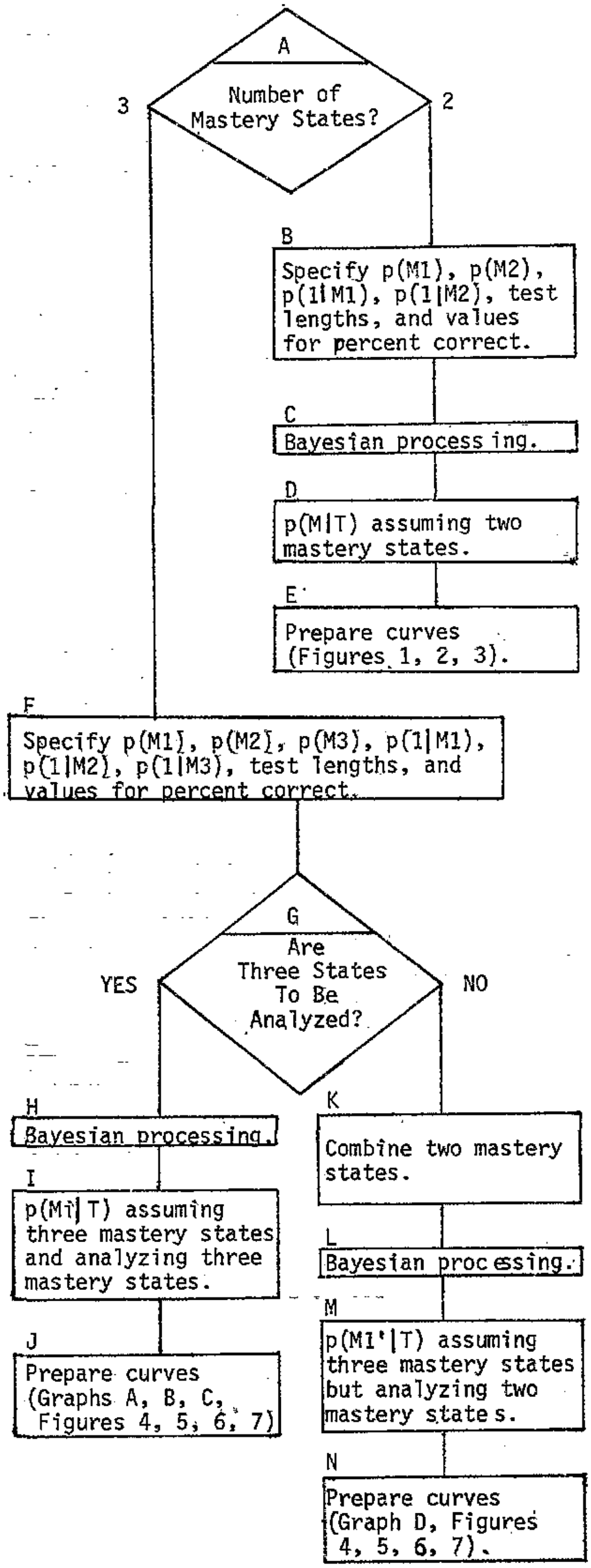
Notice that for any given test length and percent correct, the sum of the probabilities of being in states M1, M2, and M3 equals 1.0. Comparison of Graphs A, B, and C shows that when either 70% or 80% of the items for any test length are correctly answered, the probability of being in state M1 is greater than the probability of being in either state M2 or M3. That is, both the 70% and 80% curves are higher in Graph A than in either Graph B or C. For an observed score of 60% the probability of being in state M2 is greater than for M1 or M3. The probability of being in state M3 is rather low for all values of test length and percent correct observed in this particular example.

Graph D depicts the probability that a person is in mastery state M1, as opposed to a new nonmastery state composed of both M2 and M3. It can be seen that when states M2 and M3 have been thus combined, the probability of being in state M1 is greater than when all three states were analyzed independently. For observed scores of 70% or 80% correct there is slight difference in the decisions that would be made under the "independence" vs "composite" conditions. However, if a score of 60% were observed, the possibility of distinguishing between M2 and M3 would be lost when those states were combined. This loss of information may be very important if there is a large difference in cost between the selective training required for people in the M2 state and the extensive retraining needed for those in M3. This example also illustrates the potential significance of maintaining the integrity of the various nonmastery states. If the instructional decision maker knew the p(M1) with great accuracy, and also knew that there were two nonmastery states, but decided to combine the two states of nonmastery into just one state, he would be throwing away potentially valuable information. We shall return to this point in the discussion of Figure 5.

In Figure 5 the interrelationship between test length and three hypothesized mastery states becomes even more apparent. For example, Graph A shows that the probability of being in state M1 for 80% correct on a 5 item test is about 0.48. The probability of being in state M2 (shown in Graph B) for 80% correct on a five item test is about 0.36. There is thus a greater chance that a person whose score is 4 out of 5 is in M1 (p(M1|T) = 0.48), instead of M2 (p(M2|T) = 0.36) or M3 (p(M3|T) = 0.16). However, if a score of 80% correct were observed on a 40-item test, the graphs indicate that a much different decision would be appropriate. In this case, p(M1|T) equals 0.21, p(M2|T) = .78, and p(M3|T) = 0.01. Hence, people scoring 32 out of 40 correct should be classified as type M2 masters. Also note that a score of 60% for any test length implies that these people should be placed in the M3 state.

For the data used in Figure 5, the probability of finding M1 type masters is overall quite low. Instead, for the levels of achievement demonstrated by obtained scores of 60%, 70%, or 80%, it is more likely that such scores were produced by people in mastery states M2 (p(1|M2 = 0.8) and M3 (p(1|M3) = 0.6).

Graph D in this figure also represents the probability that a person is in mastery state M1 as opposed to the new (non)mastery state formed by combining states M2 and M3. In this example, all of the probabilities in Graph D are lower than in Graph A. A glance back at Figure 4, Graphs A and D reveals that the combination of states M2 and M3 increased the probability of classifying a person with a given test score as a type M1 master. Inspection of the trends in Graphs A and D of Figures 4, 5, 6, and 7 suggests that the effect of combining mastery states is to enhance the trend of the uncombined state. That is, if the probability of being in state M1 is high when the three states are treated independently, the p(M1|T) will increase after M2 and M3 are combined. Conversely, if p(M1|T) is low when the three states maintain their integrity, then combining states M2 and M3 will tend to increase the p(M1|T).


A
Number of Mastery States?
3
2
B
Specify p(M1), p(M2), p(1|M1), p(1|M2), test lengths, and values for percent correct.
C
Bayesian processing.
D
p(M|T) assuming two mastery states.
E
Prepare curves (Figures 1, 2, 3).
F
Specify p(M1), p(M2), p(M3), p(1|M1), p(1|M2), p(1|M3), test lengths, and values for percent correct.
G
Are Three States To Be Analyzed?
YES
NO
H
Bayesian processing.
I
p(Mi|T) assuming three mastery states and analyzing three mastery states.
J
Prepare curves (Graphs A, B, C, Figures 4, 5, 6, 7)
K
Combine two mastery states.
L
Bayesian processing.
M
p(M1'|T) assuming three mastery states but analyzing two mastery states.
N
Prepare curves (Graph D, Figures 4, 5, 6, 7).

## Flow-Chart Analysis of how the Bayesian Model was Developed

The impact of adding a third mastery state to the development of the model can be illustrated by tracing the logic that is required in formulating a description of the examinee population. (Refer to accompanying flow chart for a schematic summary of this discussion.) The first question the decision maker must ask (and which we considered) is: Are there two or three states of mastery inherent in the examinee population (Step A)? If two states are posited, parameter estimates for p(M1), p(M2), p(1|M1), and p(1|M2) are specified, along with plausible test lengths and values for the percent correct (Step B). The output of the Bayesian processing is the probability that a particular person is in the mastery state, p(M1|T) (Step D). A unique graph for each of Figures 1, 2, and 3 was obtained by holding the prior and conditional probabilities constant while simultaneously varying the test lengths and percent correct that would plausibly be observed (Step E). If three states are hypothesized, parameter estimates for p(M1), p(M2), p(M3), p(1|M1), p(1|M2), and p(1|M3) need to be specified, along with values for test lengths and percent correct (Step F).

Now if three states are postulated, a second decision must be made (Step G). It would seem to be usually desirable to determine the probabilities of a person's being in each of the three states (Step I). Having obtained these probabilities for selected values of prior and conditional probabilities and over a range of test lengths and percent correct scores, graphs A, B, C can be drawn such as those shown in Figures 4, 5, 6, and 7 (Step J).

However, in some instances it may be more convenient to combine the information known about two of the three mastery states. For example, even though one mastery state and two nonmastery states are hypothesized, the decision making process may require that people be divided into only two groups, of "mastery" and "nonmastery." In the present example, states M2 and M3 were combined (Step K). The result of Bayesian processing on these combined data is the probability that a person is in the new mastery state (Step M). Iteration of this procedure for various test lengths and percent correct scores over the same prior and conditional probabilities yields Graph D curves, such as those of Figures 4, 5, 6, and 7 (Step N).

The differences that result from following each of the three paths in the flow chart can be seen by comparing Figures 3A, 5A, and 5D. In each case the prior probability of being in mastery states M1 was set equal to 0.50, and the conditional probability that an M1 type master would make a correct response to an item was set equal to 0.90. Figure 3A corresponds to path A,B,C,D,E in the flow chart. Figure 5A corresponds to path A,F,G,H,I,J; and Figure 5D corresponds to path A,F,G,K,L,M,N.

In Figure 3A, p(1|M2) = 0.6, that is, a nonmaster has a 60% chance of correctly responding to an item. However, in Figure 5D the nonmastery state is the combination of states M2 and M3, with probabilities of responding correctly to an item of 0.8 and 0.6, respectively. The effect of combining M2 and M3 is to create a new (non)mastery state, where the probability of a correct response is a weighted average of the values for the uncombined groups. By defining a relatively high ability intermediate state and then combining it with a relatively low state, the probability of being in the highest mastery state is lower than if that intermediate state remained undefined. In fact, if the Figure 5 values of the prior and conditional probabilities are valid representations of the "real" states of mastery, but the values of Figure 3 (which are a simplification of the Figure 5 values) are used for decision making, then people achieving scores of 80% will be falsely classified as type M1 masters.

The differential trend between Graphs A and D of Figure 5 is noteworthy, although the absolute magnitude of the trend is rather small. For different parameter estimates (of prior and conditional probabilities), the effect of combining groups may be much more extensive. Note also that the information provided in Graph D refers only to the probability of a person's being in the mastery state, and does not directly show the loss of information about the two discrete nonmastery states that have been combined. Furthermore, when two mastery states are combined and contrasted to a third nonmastery state, the changes in the probability of being in the newly defined mastery state will often be quite different than the probability of being in one of the uncombined states.

It must be emphasized that unrealistic descriptions of the examinee population (in terms of number of mastery groups) can cause severe distortions in classification accuracy. For example, had the decision maker hypothesized only two states when, in fact, training had produced three fairly distinct states of proficiency, the results of his analysis could be highly misleading. Thus, note that the 80% line of Figure 3A ascends as more items are added (i.e., p(M1|T) increases), whereas the 80% line of

Figure 5D descends (i.e., $p(M1|T)$ decreases) as more items are added.

Caution must also be observed in the opposite case, where one might be tempted to specify more states of mastery than are actually present, in an effort to extract more information than is justified by the test data.

The present Bayesian model is not limited to three mastery states. Exploratory analyses have been conducted with up to five mastery states, and it is also hoped that the model can be generalized to deal with continuous distributions.

## Test Length and Misclassification Error

One of the most important questions that must be answered in designing a training evaluation program is: What is the probability of falsely classifying a person on the basis of a given observed score? It is also possible to turn the question around and ask: How long must a test be, and what score is required for classification decisions to be made with some specified lower limit of misclassification?

Figures 8 and 9 demonstrate how the Bayesian model can be used to answer the above questions. Assuming that the prior and conditional probabilities are realistic and fixed, the important variables are then test length and cutting score. Suppose that $p(M1) = 0.9$, $p(M2) = 0.1$, $p(1|M1) = 0.9$, and $p(1|M2) = 0.6$ as in Figure 8. In this example, the prior belief that an untested trainee is a master is very high, $p(M1) = 0.9$. A reasonable question might therefore be: What score must be observed such that a nonmastery decision can be made with at least 90% confidence? In other words, what data are required to force a reversal in the prior belief?

To be 90% confident of a nonmastery decision, $p(M2|T)$ must be equal to at least 0.90. Since the sum of $p(M1|T)$ and $p(M2|T)$ equals 1.0, $p(M1|T)$ must therefore not be greater than 0.10. Referring to Figure 8, a horizontal line crossing the

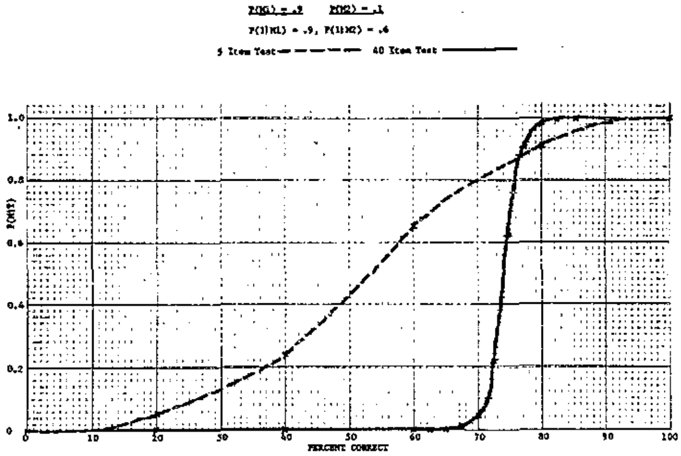


Figure 8.

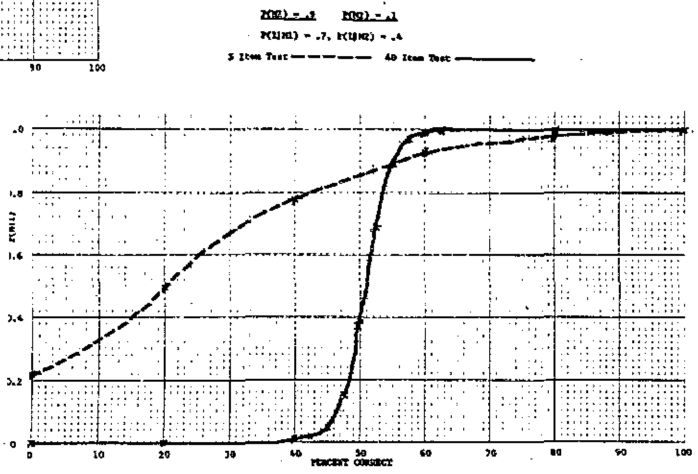


Figure 9.

ordinate at 0.10 can be drawn. This line crosses the curve for a five-item test at a point corresponding to 26% correct. The next lowest possible test score is one correct (20%), so the decision rule is that all persons scoring one correct or less should be considered nonmasters. The point on the ordinate corresponding to 20% correct on the five-item test is about 0.05. Hence, the final decision rule states that nonmastery decisions based on an observed score of one correct out of five can be made with 95% confidence (1.00 - 0.05 = 0.95). For observed scores lower than the cutoff score the confidence in making a correct decision must increase. Continuing with the present example, the p(M1|T) if zero correct are observed is virtually equal to zero. Hence, those persons who get no items right may be classified as M2 type nonmasters with nearly 100% confidence.

A similar analysis applied to the 40-item test curve indicates that the cutting score should be about 73% correct. The next lowest possible score to 73% is 70%, which yields exactly 28 correct out of 40 items. The probability of mastery given an observed score of 28 correct is about 0.04. At such a low value of p(M1|T) the chances for misclassification using a five-item test and a 40-item test are almost the same. However, the observed percent correct at which the nonmastery decision is made for the two tests is 20% on the five-item test and 70% on the 40-item test. Superficially, two tests of different lengths would seem to produce the same decision outcome, and that longer tests may not really be necessary for reducing classification error.

In order to appreciate the benefits gained by using longer length tests, the entire curve must be examined. Note that at 80% correct the five-item test yields p(M1|T) to equal 0.92. This result suggests that, on the average, 8% of the mastery decisions will be in error. For the 40-item test, the probability of mastery given 80% correct is about 0.99. That is, there is only about a 1% chance of misclassification error. A test that distinguishes sharply between masters and nonmasters is one in which the probability of mastery is close to either 0.0 or 1.00 for most obtained scores. On such tests there is only a small region in which classification error is large. For example, in Figure 8 for the 40 item test, the region where p(M1|T) is greater than 0.1 and less than 0.9 extends from 71% to 77% correct. This means that the probability of misclassifying a person will exceed 0.10 only when observed scores range from 71% to 77% correct. In contrast, the region of the five item test curve for which p(M1|T) is greater than 0.10 and less than 0.9 extends from about 26% to about 79%. Hence, there is a much larger region of the curve for which the probability of misclassification exceeds 0.10. Obviously, if classification accuracy is to be maximized over the entire range of possible test scores, then longer tests are required. Ideally, a very long test would produce a step function, for which all values of possible scores approach either 0.0 or 1.0.

Figure 9 can be analyzed in a manner similar to that for Figure 8. However, Figure 9 has one outstanding characteristic that merits special attention. If nonmastery decisions must be made with 90% confidence, and a horizontal line at p(M|T) = 0.1 is drawn, the line does not intersect the curve for the five-item test. This means that it is not possible to classify a nonmaster with 90% confidence if a five-item test is used, given the parameters used in Figure 9. If resource or time constraints are such that no more than five items may be given, and if the parameter values used in Figure 9 are realistic, and if 90% confidence for mastery decisions are required, then there is no reason to test. Testing is irrelevant because no matter what score is observed, including zero correct, the decision rule compels a mastery decision to be made. In fact, for the present values, the probability of mastery given zero correct, is equal to 0.21. This simply means that if persons obtaining a score of zero are classified as nonmasters, 21% of them will be misclassified, on the average.

The implication of these results for performance testing is obvious. Since performance tests are often rather short, it is essential that the magnitude of misclassification error that can be incurred with such tests be recognized. Designing tests that have clear and direct relation to actual performance is certainly a worthwhile and much needed effort. However, reasonable levels of confidence in classifying trainees must not be sacrificed merely for the sake of using conveniently short tests.

## Appendix I: A Computational Example for Three Mastery States

The following example illustrates the computations necessary for processing data with the Bayesian model. The values chosen for this example correspond to Figure 4. Assume that there are three states of mastery, and unequal prior probabilities for these three states. The educational decision-maker must provide estimates for the prior probabilities of master, p(Mi). For this example let us assume the values to be: p(M1) = .5; p(M2) = .3; and p(M3) = .2. He must also provide estimates for the conditional probability of getting any given test item right, given each mastery state. The following values will be used as the conditional probability of getting an item right given a mastery state: p(1|M1) = .8; p(1|M2) = .6; p(1|M3) = .5. The conditional probabilities of getting an item wrong given a mastery state are: p(0|M1) = .2; p(0|M2) = .4; and p(0|M3) = .5.

First we need to calculate the probability that an item is answered correctly. For the overall population, p(tj = correct)

$$= \sum_{i=1}^{S} p(Mi)p(tj = \text{correct}|Mi) = (.5)(.8) + (.3)(.6) + (.2)(.5) = .68.$$

Likewise,

$$p(tj = \text{wrong}) = \sum_{i=1}^{S} p(Mi)p(tj = \text{wrong}|Mi = (.5)(.2) + (.3)(.4) + (.2)(.5) = .32.$$

We also need to obtain the set of conditional probabilities for the different mastery states given than an individual item was responded to either correctly or wrongly. The general equation is:

$$p(Mi|tj) = \frac{p(Mi)p(tj|Mi)}{p(tj)}.$$

Substituting the above values yields:
p(M1|tj = correct) = (.5)(.8) ÷ .68 = .588;
p(M2|tj = correct) = (.3)(.6) ÷ .68 = .265;
and p(M3|tj = correct) = (.2)(.5) ÷ .68 = .147.
(Note that the sum equals 1.0.) Finally,
p(M1|tj = wrong) = (.5)(.2) ÷ .32 = .3125
p(M2|tj = wrong) = (.3)(.4) ÷ .32 = .375 and
p(M3|tj = wrong) = (.2)(.5) ÷ .32 = .3125

If 6 items were answered correctly on a 10-item criterion-referenced test, the following $\prod_{j=1}^{N} p(Mi|tj)$ values result:
M1 = $3.9 \times 10^{-4}$; M2 = $6.8 \times 10^{-6}$;
M3 = $9.6 \times 10^{-8}$

Finally, the general Bayesian formula yields the conditional probability for each mastery state given the total test score. For example, p(Mi|T) =

$$\frac{(3.9 \times 10^{-4})}{(.5)^9 \left[\frac{(3.9 \times 10^{-4})}{(.5)^9} + \frac{(6.8 \times 10^{-6})}{(.3)^9} + \frac{(9.6 \times 10^{-8})}{(.2)^9}\right]}$$

= .272.

Similar calculations yield p(M2|T) = .473 and p(M3|T) = .254.

In order to combine mastery states M2 and M3 into a single mastery state (which could represent combining the two degrees of nonmastery, Figure 4, Graph D), the following calculations are required. The values for p(M1) and $\prod_{j=1}^{N} p(M1|tj)$ remain the same, .5 and 3.9 x $10^{-4}$ respectively. The new nonmastery state (M2') occurs as a result of combining the previous states M2 and M3. Hence, p(M2') = p(M2) + p(M3) = .3 + .2 = .5, p(M2'|tj = correct) = p(M2|tj = correct) + p(M3|tj = correct) = .265 + .147 = .412, and p(M2'|tj = wrong) = p(M2|tj = wrong) + p(M3|tj = wrong) = .375 + .3125 = .6875. Calculation of $\prod_{j=1}^{N} p(M2'|tj)$ yields $1.09 \times 10^{-3}$.

Entering these new values into the general Bayesian Formula, the following values of p(M1'|T) and p(M2'|T) are obtained:

$$p(M1'|T) = \frac{3.9 \times 10^{-4}}{(.5)^9\left[\frac{(3.9 \times 10^{-4})}{(.5)^9} + \frac{(1.09 \times 10^{-3})}{(.5)^9}\right]}$$

= .264,

$$p(M2'|T) = \frac{1.09 \times 10^{-3}}{(.5)^9\left[\frac{(3.9 \times 10^{-4})}{(.5)^9} + \frac{(1.09 \times 10^{-3})}{(.5)^9}\right]}$$

=.736.

Some interesting properties of the model emerge when an alternative procedure for combining mastery groups is used. Notice that in order to combine two mastery states it is not necessary to calculate new values for p(1|M2') and p(0|M2'). However, it is possible to show that these values are weighted averages of p(1|M2) and p(1|M3), and p(0|M2) and p(0|M3) respectively where the weights are the relative proportions of the new state accounted for by each of the previous states. The calculations follow.

Since p(M2) = .3 and p(M3) = .2, state M2 accounts for 60% and M3 accounts for 40% of the new state M2'. Hence, the value of p(1|M2') = (.6)p(1|M2) + (.4)p(1|M3) = (.6)(.6) + (.4)(.5) = .56 and p(0|M2') = (.6)p(0|M2) + (.4)p(0|M3) = (.6)(.4) + (.4)(.5) = .44.

Using these new values, $p(tj = correct) = p(M1')p(1|M1') + p(M2')p(1|M2') = (.5)(.8) + (.5)(.56) = .68$ and $p(tj = wrong) = p(M1')p(0|M1') + p(M2')p(0|M2') = (.5)(.2) + (.5)(.44) = .32$.

Finally, $p(M2'|1)$ and $p(M2'|0)$ may be calculated.

$$p(M2'|1) = \frac{p(M2')\ p(1|M2')}{p(1)} = \frac{(.5)(.56)}{.68} = .412,$$

and

$$p(M2'|0) = \frac{p(M2')\ p(0|M2')}{p(0)} = \frac{(.5)(.44)}{.32} = .6875.$$

These values are the same as those obtained by the simple addition procedure shown above.

This exercise serves to illustrate the effect of combining two mastery states. Combining states M2 and M3, creates, in effect, a new description of the examinee population in which only two mastery states are hypothesized. The parameter estimates for the new states in this example, are

$$p(M1) = .5 \qquad p(M2) = .5$$
$$p(1|M1) = .8 \qquad p(1|M2) = .56.$$

In choosing to combine groups, the decision maker must consider whether a two state description of the population with parameter estimates such as those above is a better representation than the original three state descriptions with parameter estimates

$$p(M1) = .5,\ p(M2) = .3,\ p(M3) = .2,$$
$$p(1|M1) = .8,\ p(1|M2) = .6,\ p(1|M3) = .5.$$

## References

Hershman, R.L. A rule for the integration of Bayesian opinions. Human Factors, 1971, 13, 255-259.

Novick, M.R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Center for the Study of Evaluation Monograph Series in Evaluation, III: Problems in criterion-referenced measurement. Los Angeles: U.C.L.A. Center for the Study of Evaluation, 1974.

The views expressed in this paper are those of the authors and do not imply endorsement by the U. S. Army.

## *ABOUT THE AUTHORS*

*MR. KENNETH I. EPSTEIN is with the U.S. Army Research Institute (ARI) for the Behavioral and Social Sciences. Since joining ARI, he has been primarily concerned with the development of a theoretical base to support training evaluation, with particular emphasis on the application of statistical and psychometric theory to criterion-referenced measurement problems. He received an M.S. degree in Educational Research and Testing from Florida State University, where he participated in the development of the Instructional Systems Development Model for the U.S. Army Combat Arms Training Board and the Interservice Committee on Instructional Systems Development. He is currently completing the requirements for a Ph.D. His research area is the application of criterion-referenced measurement to instructional decision making.*

*DR. FREDERICK H. STEINHEISER, JR. is with the U.S. Army Research Institute for the Behavioral and Social Sciences, where his main research interests are in decision-making aspects of training and trainee evaluation, as well as the study of the multidimensional components of judgment in setting standards for acceptable trainee and group performance. He received a B.A. degree from the University of Michigan and a Ph.D. from the University of Cincinnati, both in Human Experimental Psychology. He then spent 2 years as an N.I.M.H. Postdoctoral Fellow at the Center for Research in Human Learning of the University of Minnesota studying language and communications. This was followed by a Research Associateship at the Johns Hopkins University, for the study of reading disability using information-processing theory and methodology. Industrial experience includes the production of a semi-programmed medical electronics manual and consultantship to devise methods for promoting energy conservation.*