Applications of Sequential Testing Procedures to Performance Testing

Kenneth I. Epstein
Claramae S. Knerr
U.S. Army Research Institute
for the Behavioral and Social Sciences

Methods for accurately and efficiently making classification decisions are becoming more necessary in applications of educational and psychological testing. The systematic design of instruction requires that data indicating the degree of effectiveness of instructional materials be collected and used to identify that which is acceptable and that which needs improvement. Criterion-referenced testing implies that students will be classified on the basis of whether or not they have mastered specified instructional objectives. Performance testing usually refers to a special case of criterion-referenced testing that applies to training programs emphasizing job-related skills. Trainees are classified on the basis of performance test scores as sufficiently skilled or inadequately trained.

Typically, when instructional effectiveness is evaluated, the classification decision is dependent upon whether or not some criterion proportion of students successfully completes a program of instruction. For the criterion-referenced or performance testing case, classification depends upon whether or not the criterion proportion of test items is passed. This paper is concerned with two problems associated with these procedures: (1) the number of students or items tested and (2) the specification of misclassification error rates. Before addressing these problems directly, it will be necessary to clarify the testing problem and to state several important assumptions concerning the nature of the data.

The Classification Problem in Testing

Baker and Alkin (1973) have suggested that one of the critical factors in judging instructional effectiveness is the extent to which learners master the objectives. This can be taken one step further to conceptualize instructional effectiveness in terms of the extent to which *any* student in the target population is likely to master the objectives, given the opportunity. This is exactly what is implied in the 80/80 criterion often applied to instructional development efforts; the instruction is considered effective if at least 80% of the students who begin the instruction complete at least 80% of the objectives on the first attempt.

Criteria of Instructional Effectiveness

Let it be assumed, for the moment, that the 80/80 criterion is reasonable for a particular instructional development effort. That is, although all of the students who begin the instruction may not successfully complete 80% of the objectives on the first attempt, if 80% of them do, it will be satisfactory. Let it also be assumed that valid and reliable measures of the objectives are available. How is it determined whether or not the instruction is acceptable in its present state? The obvious answer is to find subjects who are members of the population for which the instruction is intended, to let them try the instructional materials, and to find out how well they perform on a test of the objectives.

Under the 80% criterion rule, the instruction will be considered effective if at least 80% of the students in the target population accomplish the objective. This criterion may also be interpreted as the probability that any randomly chosen student will accomplish the objective, that is, .80. In other words, the performance of a randomly chosen student may be considered a Bernoulli variable, with the probability of success equal to instructional effectiveness. Either an acceptance or rejection decision can be made when sufficient evidence to draw inferences about the instructional effectiveness has been gathered.

Sample Size

This leads directly to the question of the size of the tryout sample of students. The data gathered during a tryout of the instruction which is designed to meet this criterion is used to draw inferences about the effectiveness of the instruction for the total target population. Clearly, the larger the sample of students in the tryout group, the better the estimate of instructional effectiveness for the total group will be.

An indication of the precision with which population parameters are estimated by sample data is given by confidence limits. For example, assume that the tryout sample for one objective consists of five students, four of whom accomplish the objective. The effectiveness of the instruction, in terms of the proportion of students who accomplished the objective, is 80%. However, the 95% confidence limits for a proportion based on four correct answers in five trials are .343 and .990. These confidence limits assume that a random sample is drawn from an infinitely large population. Since most instructional development efforts involve materials which will be useful for a large number of students and since students for a tryout should be randomly sampled from the target population, this assumption seems reasonable.

The relatively widely separated values of the confidence limits imply that extreme caution should be taken in drawing any inferences about the instructional effectiveness for the total population from a tryout sample of five. Unfortunately, increasing the sample size, while staying within the bounds of practical constraints, is not very helpful. For example, the 95% confidence limits for a proportion based on observing 8 correct in 10 trials are .397 and .963; for 16 correct in 20 trials, .589 and .929; for 24 correct

in 30 trials, .636 and .909; for 40 correct in 50 trials, approximately .67 and .90; and for 80 correct in 100 trials, approximately .71 and .88. Novick and Lewis (1974) discuss the sample size problem in terms of errors in decision making from a criterion-referenced point of view. Their conclusions are equally discouraging for small samples of data.

Misclassification

The misclassification question is also difficult, partly because it is rarely addressed or understood in practical applications of instructional materials development. The first problem is choosing a criterion proportion of students in the target population for whom the instruction should be effective. Ideally, all students should be reached; but this is an unrealistic goal. Individual student differences, costs involved in designing instructional materials, and limitations on instructional time and resources all mitigate against achieving this ideal. Consumers and developers of instructional materials must begin to consider these factors and set reasonable goals for themselves based both on their specific needs and on the student populations. While the sequential testing procedure to be described in this paper does not offer a method for determining the criterion, it does demand that a criterion be specified. Perhaps by requiring such specificity, discussions necessary to arrive at meaningful criteria will transpire.

A second requirement of the procedure is that the decision maker realize that since he/she is only able to sample part of the target population, occasionally incorrect decisions will be made. Thus, very good instruction will sometimes be rejected, based on tryout results--a false negative decision--and poor instruction will sometimes be accepted--a false positive decision. The costs of such misclassifications must be considered in deriving a decision-making procedure. If misclassifications are considered extremely costly, then a larger tryout sample must be used. If the costs of false positive and false negative decisions are not equal, then adjustments must be made in the decision rule to insure that the likelihood of the more costly incorrect decisions is kept small. As in the case of the criterion associated with acceptable instruction, the sequential testing procedure does not solve the misclassification rate problem. The procedure does, however, force decision makers to come to grips with the problem.

Criterion-Referenced and Performance Testing

The same types of problems faced by instructional materials designers are encoutered by criterion-referenced and performance test developers. The test is typically a relatively small sample of items or problem situations drawn from a large pool defined by the objectives. Time and resource constraints require that the number of items be kept small. Mastery of the objective is stated as a criterion proportion of correct responses that would be obtained if all items in the pool could be administered. A decision rule is required that classifies students as masters or non-masters of an objective, based on their responses to the sample of items included on the test.

The test designer must answer questions analogous to those of the instructional designer: (1) What should the criterion proportion correct be?

(2) How many items should be included on the test? (3) What are the relative costs of false positive and false negative decisions concerning student ability? (4) How can they be quantified? and (5) How can these factors be incorporated in a usable decision rule?

Sequential Analysis in Instructional Design and Criterion-Referenced Testing

The following assumptions apply to the data collected both for instructional materials evaluation and student evaluation. The data are a series of dichotomous pass-fail decisions. In each case, the sample proportions of pass-fail decisions represent estimates of those proportions that would be obtained if all students in the target population or all items in an objective's domain could be tested. Each student or item is considered an independent random sample from the population and has associated with it a probability of passing equal to the population proportion of passing decisions. The particular sequence of students tested or items presented does not represent a systematic order effect. Each individual student or item pass-fail decision is unambiguous.

Sequential analysis was developed as an alternative to traditional statistical hypothesis testing by Wald in the early 1940s. The theory and techniques are described in Wald's Sequential Analysis.[1] This paper will be concerned only with the application of the general theory to testing the mean of a binomial distribution (Wald, 1973, pp. 88-105); the mathematics involved will be described and the necessary parameters in terms of instructional effectiveness and criterion-referenced testing will be defined. This will be followed by a computed example and the results of three applications of the procedure.

Procedures of Sequential Analysis

Wald's context for describing the procedure was acceptance inspection of a lot consisting of a large number of manufactured products. The problem was to determine whether or not the proportion of defective items exceeded some predetermined limit based on the inspection of a relatively small sample of items. Similarly, the instructional design and criterion-referenced testing problems are to determine whether or not the criterion proportions required for acceptance were exceeded by (1) the proportion of students for whom the instruction was ineffective and (2) the proportion of items answered incorrectly on a criterion-referenced test.

Accept/Reject decisions. Let p equal the population proportion of students for whom instruction is ineffective or the proportion of items an individual student would answer incorrectly, given all the items in the domain. Let p_c equal the corresponding criterion proportions required for acceptance. If $p<p_c$, the correct decision is to accept. If $p>p_c$, the correct decision is to reject. Since decisions will be based on sample rather than population

[1] Page references in this paper refer to the unabridged and unaltered republication of the 1947 work by Dover Publications, Inc., 1973.

data, misclassification errors must be considered. When $p<p_c$, the preference to accept increases with decreasing values of p. When $p>p_c$, the preference to reject increases with increasing values of p. However, since errors will occur, it should be possible to define some p_o slightly below p_c where an incorrect rejection decision has little practical consequence. Similarly, some p_1 slightly above p_c can be chosen where an incorrect acceptance decision is not serious. The region $p_o<p_c<p_1$ is known as an indifference region. If the population proportion, p, falls within the indifference region, the practical consequences of an incorrect decision are negligible.

Once the limits of the indifference region have been specified, it becomes reasonable to choose values for the risks of incorrect decisions. The probability of a rejection decision, when $p\leq p_o$, should not exceed some small value α; and the probability of an acceptance decision, when $p\geq p_1$, should not exceed some small value β. In other words, α is the acceptable risk of committing a false negative error and β is the acceptable risk of committing a false positive error.

Given values for the following five parameters, a sequential sampling plan can be specified:

1. p_c, the population proportion or probability of failures or incorrect responses that defines an unacceptable product or student,
2. $p_o < p_c$, a lower limit proportion or probability of failures or incorrect responses below which false negative errors are critical;
3. $p_1 > p_c$, an upper limit proportion or probability of failures or incorrect responses above which false positive errors are critical;
4. α, the acceptable risk of committing a false negative error; and
5. β, the acceptable risk of committing a false positive error.

The choice of the parameter values is not a statistical problem. Rather, the values must be chosen on the basis of the practical considerations and requirements of each particular test.

Sequential sampling. The sequential sampling procedure takes advantage of the need for relatively little data to identify a very good or a very poor product or student and the requirement of extensive observations only for marginal cases. For each observation, the probabilities of the observation and all preceding observations, given $p=p_o$ and $p=p_1$, are calculated and the log of their ratio obtained. An accept, reject, or continue sampling decision is made based on the value of the log probability ratio. Wald has shown that an accept or reject decision will eventually occur with probability 1.0 and that fewer observations are required, on the average, for given values of p_o, p_1, α, and β using the sequential sampling procedure than are required by traditional hypothesis testing procedures.

Let m equal the number of misses or failures in the first n observations. Given $p=p_o$, the probability of observing such a sample is

$$p_{om} = p_o^{\,m}(1-p_o)^{n-m} \quad . \qquad [1]$$

For $p=p_1$, the probability is

$$p_{1m} = p_1^{\,m}(1-p_1)^{n-m} \quad . \qquad [2]$$

The log of the ratio is now computed

$$\log \frac{p_{1m}}{p_{om}} = \log \frac{p_1^{\,m}(1-p_1)^{n-m}}{p_o^{\,m}(1-p_o)^{n-m}} = m \log \frac{p_1}{p_o} + (n-m) \log \frac{(1-p_1)}{(1-p_o)} \quad . \qquad [3]$$

After each observation, the following inequality is evaluated:

$$\log \beta < \log \frac{p_{1m}}{p_{om}} < \log A \quad . \qquad [4]$$

If $\log \frac{p_{1m}}{p_{om}} > \log A$, then the test terminates with a rejection decision.

If $\log \frac{p_{1m}}{p_{om}} < \log B$, then the test terminates with an acceptance decision.

Otherwise, another observation is made and the decision rule is applied again.

The values of A and B are functions of α and β. Wald (1973, pp. 41-48) discusses the relationships between A, B, α, and β and shows that very close approximations to the exact values required for hypothesis testing are obtained by setting

$$A = [(1-\beta)/\alpha] \qquad [5]$$

and

$$B = [\beta/(1-\alpha)] \quad . \qquad [6]$$

In fact, $[(1-\beta)/\alpha]$ and $[\beta/(1-\alpha)]$ are upper and lower limits for the exact values of A and B, respectively, which implies that the use of the approximations will provide a slightly conservative test.

Replacing A and B with their approximations and performing several algebraic steps leads to a convenient form of the decision-making inequality:

$$\log[\beta/(1-\alpha)] < m \log \frac{p_1}{p_o} + (n-m) \log \frac{(1-p_1)}{(1-p_o)} < \log [(1-\beta)/\alpha] \qquad [7]$$

$$\log[\beta/(1-\alpha)] < m \left(\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)} \right) + n \log \frac{(1-p_1)}{(1-p_o)} < \log \frac{1-\beta}{\alpha} \qquad [8]$$

$$\frac{\log \frac{\beta}{1-\alpha} - n \log \frac{(1-p_1)}{(1-p_o)}}{\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)}} < m < \frac{\log \frac{1-\beta}{\alpha} - n \log \frac{(1-p_1)}{(1-p_o)}}{\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)}} \quad . \qquad [9]$$

Using this form of the equation, the number of misses (m) is compared to the extremes of the inequality. If m is less than or equal to the left hand extreme, the test terminates with an acceptance decision. If m is greater than or equal to the right hand extreme, the test terminates with a rejection decision; otherwise, sampling continues. An interactive computer scoring scheme would simplify this process.

Graphical procedures. If a computer is not available, a graphic form of the procedure can be used. The abscissa represents the number of observations; the ordinate represents the number of misses. The extremes of the inequality represent two parallel lines on the graph with common slope:

$$\frac{-\log \frac{(1-p_1)}{(1-p_o)}}{\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)}} = s \qquad [10]$$

and intercepts

$$\frac{\log \frac{\beta}{1-\alpha}}{\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)}} = I_a \qquad [11]$$

and

$$\frac{\log \frac{1-\beta}{\alpha}}{\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)}} = I_r \qquad [12]$$

for the left and right extremes respectively. Each datum is plotted as it is observed. As soon as the graph of the observations crosses the upper line, the test terminates with a rejection decision. As soon as it crosses the lower line, the test terminates with an acceptance decision; otherwise, the sampling continues. Using s, I_r, and I_a to represent the common slope, the intercept of the rejection line, and the intercept of the acceptance line, the decision rule can be simply stated:

1. Reject if $m \geq I_r + ns$;
2. Accept if $m \leq I_a + ns$;

3. Continue sampling if $I_a + ns < m < I_r + ns$.

Several other useful functions are also easy to calculate. The minimum number of observations--all misses--required for a rejection decision is the next higher integer greater than $[I_r/(1-s)]$. The minimum number of observations--all passes--required for an acceptance decision is the next higher integer greater than $(I_a/-s)$. The operating characteristic (OC) function of the test, $L(p)$, is the probability of an acceptance decision for p, the population proportion of misses. It is usually only necessary to calculate five values of $L(p)$ to graph the OC function, since the general shape of the curve is similar for any sequential sampling plan.

Five convenient values are

1. $L(o) = 1$; for $p=o$ the test will always terminate eventually with an acceptance decision
2. $L(1) = 0$; for $p=1$ the test will always terminate eventually with a rejection decision
3. $L(p_o) = 1-\alpha$, by definition
4. $L(p_1) = \beta$, by definition
5. $$L(p=s) = \frac{\log \frac{1-\beta}{\alpha}}{\log \frac{1-\beta}{\alpha} + \left|\log \frac{\beta}{1-\alpha}\right|} \qquad [13]$$

$$= \frac{I_r}{I_r + |I_a|}$$ (Wald, 1973, p. 95).

Wald (1973, pp. 96-98) provides a general equation for any value of $L(p)$:

$$L(p) = \frac{\left(\frac{1-\beta}{\alpha}\right)^h - 1}{\left(\frac{1-\beta}{\alpha}\right)^h - \left(\frac{\beta}{1-\alpha}\right)^h}, \qquad [14]$$

where h can take any value between $\pm \infty$ and

$$p = \frac{1-\left(\frac{1-p_1}{1-p_o}\right)^h}{\left(\frac{p_1}{p_o}\right)^h - \left(\frac{1-p_1}{1-p_o}\right)^h}. \qquad [15]$$

Finally, the expected number of observations, $E_p(n)$, necessary to reach a decision can be calculated as a function of p (Wald, 1973, pp. 99-101). The general equation is

$$E_p(n) = \frac{L(p) \log \frac{\beta}{1-\alpha} + [1-L(p)] \log \frac{1-\beta}{\alpha}}{p \log \frac{p_1}{p_o} + (1-p) \log \frac{1-p_1}{1-p_o}} \qquad [16]$$

Special formulas for the values of p used to calculate the OC function are

$$E_{p=o}(n) = \frac{\log \frac{\beta}{1-\alpha}}{\log \frac{1-p_1}{1-p_o}} \qquad [17]$$

$$E_{p=1}(n) = \frac{\log \frac{1-\beta}{\alpha}}{\log \frac{p_1}{p_o}} \qquad [18]$$

$$E_{p=p_o}(n) = \frac{(1-\alpha) \log \frac{\beta}{1-\alpha} + \alpha \log \frac{1-\beta}{\alpha}}{p_o \log \frac{p_1}{p_o} + (1-p_o) \log \left(\frac{1-p_1}{1-p_o}\right)} \qquad [19]$$

$$E_{p=p_1}(n) = \frac{\beta \log \frac{\beta}{1-\alpha} + (1-\beta) \log \frac{1-\beta}{\alpha}}{p_1 \log \frac{p_1}{p_o} + (1-p_1) \log \left(\frac{1-p_1}{1-p_o}\right)} \qquad [20]$$

$$E_{p=s}(n) = \frac{-\log \frac{\beta}{1-\alpha} \log \frac{1-\beta}{\alpha}}{\log \frac{p_1}{p_o} \log \left(\frac{1-p_o}{1-p_1}\right)} \qquad [21]$$

where $E_{p=s}(n)$ is or very nearly approximates the maximum value of the function.

Numerical Example

The three applications of the procedure (discussions of which follow) used the same values for the necessary parameters. Those values were used in a computational example and will hold for the remainder of the paper.

1. The criterion probability of a student not passing minimally acceptable instruction materials or the criterion probability of a student not correctly answering a criterion-referenced test item, p, is .20.
2. The probability of a student not passing instruction or a student incorrectly answering a test item below which false negative errors are critical, p_o, is .10.
3. The probability of a student not passing instruction or a student incorrectly answering a test item above which false positive errors are critical, p_1, is .30. Thus, the indifference region will be $.10 < p < .30$.
4. The risk of committing a false negative error, α, is .01.
5. The risk of committing a false positive error, β, is .10.

Given these parameter values, the decision-making inequality can be computed:

$$s = \frac{-\log \frac{.70}{.90}}{\log \frac{.30}{.10} - \log \frac{.70}{.90}} = \frac{-\log .778}{\log 3 - \log .778} = .186$$

$$I_a = \frac{\log \frac{.10}{.99}}{\log \frac{.30}{.10} - \log \frac{.70}{.90}} = \frac{\log .101}{\log 3 - \log .778} = -1.70$$

$$I_r = \frac{\log \frac{.90}{.01}}{\log \frac{.30}{.10} - \log \frac{.70}{.90}} = \frac{\log 90}{\log 3 - \log .778} = 3.33.$$

Therefore, continue to sample if

$$-1.70 + .186\, n < m < 3.33 + .186\, n,$$

where n is the total number of observations and m is the number of failures or incorrectly answered items.

Figure 1 is the graph of the decision-making inequality. Since $[I_r/(1-s)] = 3.33/.814 - 4.09$, the minimum number of observations necessary

Figure 1
Sequential Decision Making Graph

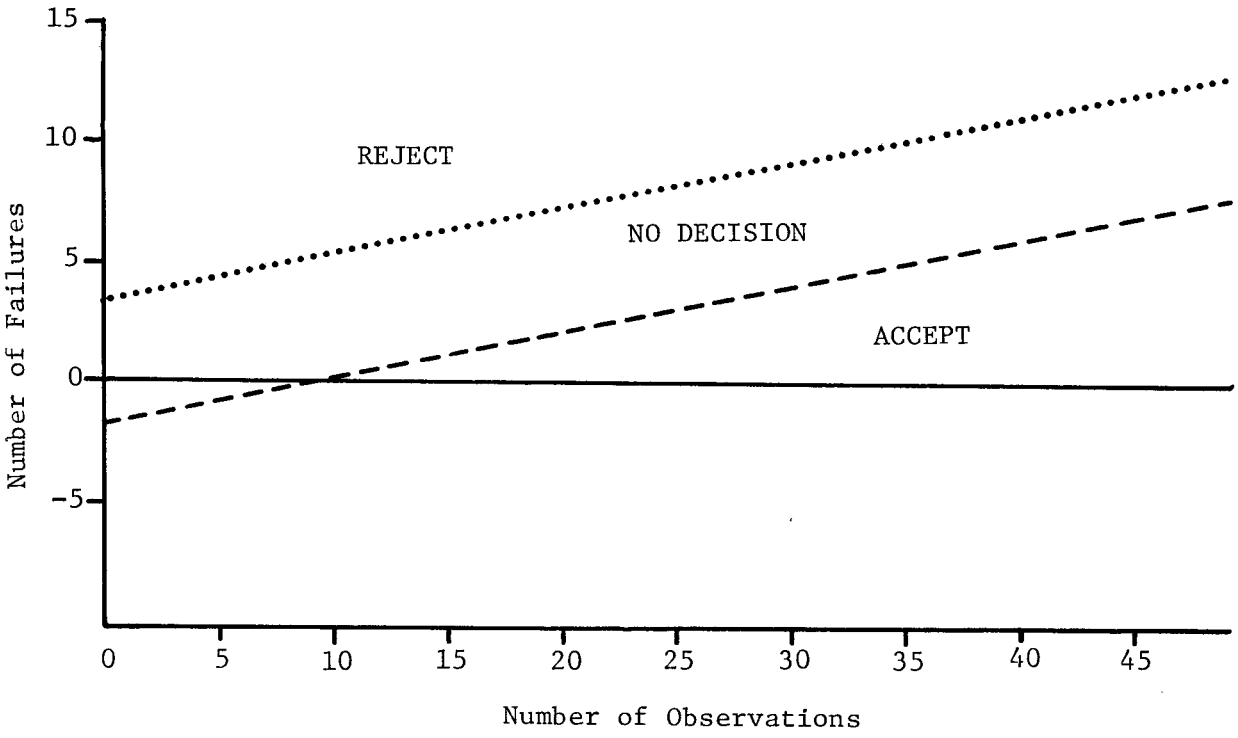
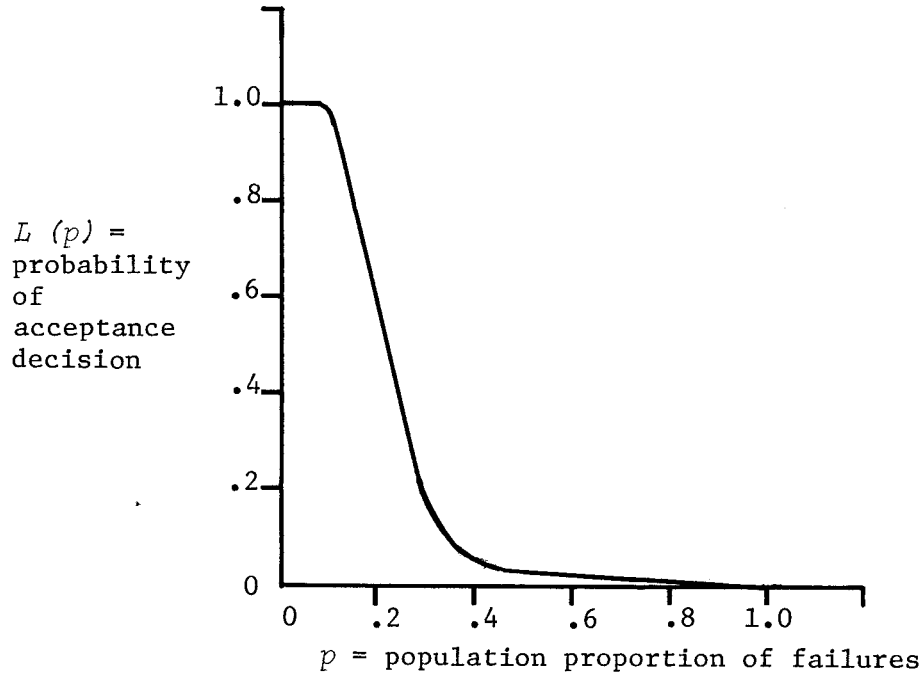


Figure 2
Sequential Testing Operating Characteristic Function



to reject will be five. Furthermore, since $Ia/-s = -1.70/.186 = 9.14$, the minimum number of observations necessary to accept will be 10.

A graph of the OC function is shown in Figure 2 using the following values:

1. $L(p=o) = 1$
2. $L(p=1) = 0$
3. $L(p=p_o=.10) = 1 - .01 = .99$
4. $L(p=p_1=.30) = .10$
5. $L(p=s= .186) = \frac{3.33}{3.33 + 1.70} = .66.$

Figure 3 is a graph of the expected number of observations using the following values:

1. $$E_{p=o}(n) = \frac{\log \frac{.10}{.99}}{\log \frac{.70}{.90}} = \frac{\log .101}{\log .778} = 9.14$$

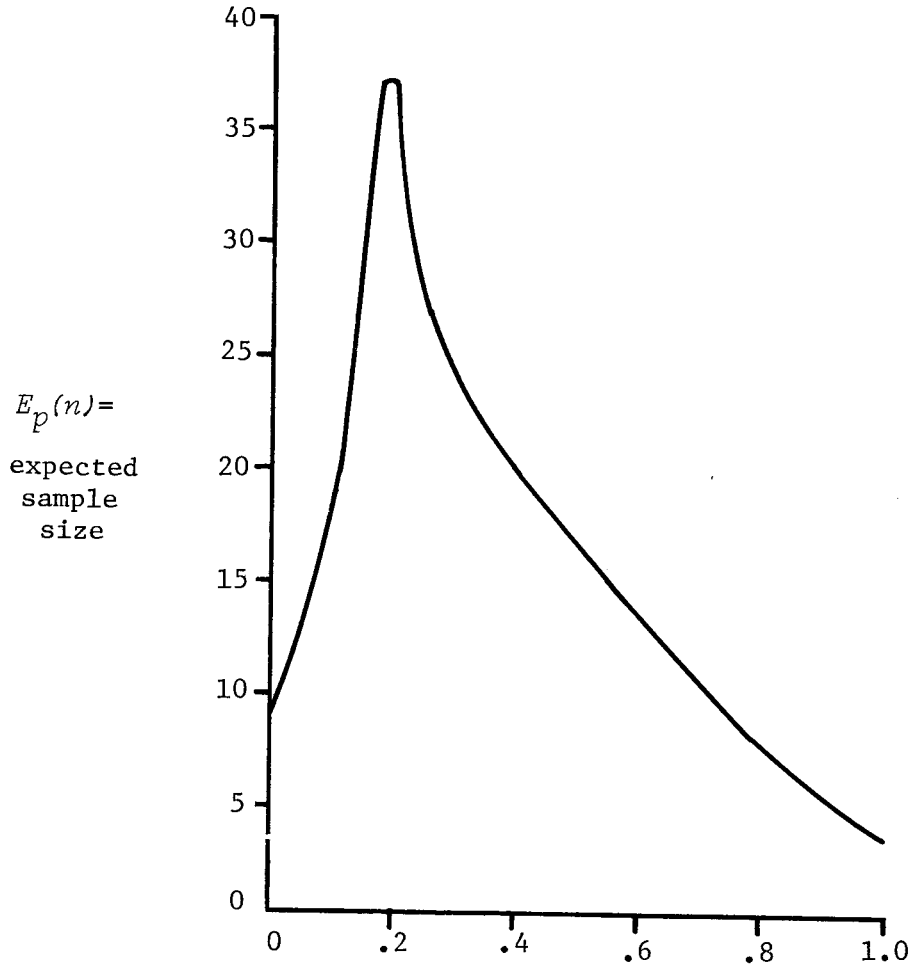
2. $$E_{p=1}(n) = \frac{\log \frac{.90}{.01}}{\log \frac{.30}{.10}} = 4.09$$

3. $$E_{p=p_o=.10}(n) = \frac{.99 \log \frac{.10}{.99} + .01 \log \frac{.90}{.01}}{.10 \log \frac{.30}{.10} + .90 \log \frac{.70}{.90}}$$
$$= \frac{.99 \log .101 + .01 \log 90}{.10 \log 3 + .90 \log .778} = 18.58$$

4. $$E_{p=p_1=.30}(n) = \frac{.10 \log \frac{.10}{.99} + .90 \log \frac{.90}{.01}}{.30 \log \frac{.30}{.10} + .70 \log \frac{.7}{.9}}$$
$$= \frac{.10 \log .101 + .90 \log 90}{.30 \log 3 + .70 \log .778} = 24.84$$

5. $$E_{p=s=.186} = \frac{-\log \frac{.10}{.99} \log \frac{.90}{.01}}{\log \frac{.30}{.10} \log \frac{.90}{.70}}$$
$$= \frac{-\log .101 \log 90}{\log 3 \log 1.29} = 37.43 .$$

Figure 3
Expected Sample Size for the Sequential Test



Applications

Instructional Materials Data

The first set of example data is a re-analysis of some instructional materials tryout data (Epstein, 1975). It is supportive of the validity of sequential decisions and clearly illustrates the savings possible in the size of the student tryout pool.

Data. The U.S. Army has been heavily involved in the design of audio-visual instruction to teach a wide variety of skills. Tryout data were available for instruction to teach land navigation. The instruction covered eight objectives, each objective having associated with it a performance test that was scored pass/fail. Twenty-eight students participated in the tryout. The data are shown in Table 1.

Table 1
Results of U.S. Army Tryout for Audio Visual Instruction in Land Navigation, n=28

Objective Number	Number Passing	Percent Passing	95% Confidence Limits for Proportion	
1	27	96	.830	.998
2	26	93	.783	.987
3	26	93	.783	.987
4	25	89	.741	.970
5	26	93	.783	.987
6	26	93	.783	.987
7	20	71	.537	.858
8	16	57	.381	.742

The decision rule used to evaluate this instruction was the 80/80 rule that 80% of the students pass 80% of the objectives. The 95% confidence limits imply that the instruction was certainly acceptable for Objective 1 and that relatively high confidence can be placed in the effectiveness of the instruction for Objectives 2, 3, 4, 5, and 6. The effectiveness of Objective 7 is questionable, and the instruction for Objective 8 is certainly below the minimum requirement.

Results. A sequential testing strategy was applied to the same data. The values for p_o, p_1, α, and β were the same values used in the example discussed earlier. The results of the sequential testing procedure are summarized in Table 2. Since there was no reason to believe that the students were arranged in any particular order, the results from Student 1 in the Army tryout were plotted first, the results from Student 2 second, and so forth.

Table 2
Results of Sequential Testing for Tryout Data from Army Audio-Visual Instruction in Land Navigation

Objective Number	Number Tested	Decision
1	10	Accept
2	15	Accept
3	15	Accept
4	20	Accept
5	10	Accept
6	20	Accept
7	17	Reject
8	6	Reject

Average Number Tested = 14.125

The results of the sequential testing procedure agree with the results obtained using the 80/80 rule with the 28 subjects; that is, the instruction was accepted for Objectives 1 through 6 and revised for Objectives 7 and 8. In all cases, fewer students were needed than when using the 80/80 rule. In fact, about half as many students were tested, on the average; and the results for Objective 8 (clearly the objective for which revisions were most needed) were obtained with only six students.

Mastery Decisions

Data. The second set of example data is an application of sequential testing for making individual mastery/nonmastery decisions. It is also a re-analysis of existing data. A total of 237 Military Police students were tested, using the .45 caliber handgun to fire at stationary silhouette targets. Each student fired a total of 240 shots (trials) over a period of two days.

The 240 trials were divided into 3 repetitions of 80 trials each. The first repetition was fired the morning of the first day, the second that afternoon, and the third the following morning. Each group of 80 was subdivided into 10 shots for each of 8 "tables," or distance-position combinations, which define the standard operating procedure for the Military Police Firearms Qualification Course (MPFQC). For this test each group of 10 shots was subdivided into 2 groups of 5 shots. The student had to reload after taking the first 5 shots before shooting the second group of 5 for the same table. The firing order and a description of each table are shown in Table 3.

Table 3
The Military Police Firearms Qualification Course (MPFQC)

Table	Range (meters)	Position	Maximum Time (min.: sec.)
1	35	lying prone	1:45
2	25	standing, no support, preferred hand	1:45
3	25	standing with support, weak hand	1:30
4	25	standing with support, preferred hand	1:30
5	15	standing, no support preferred hand	1:20
6	15	kneeling with support, left hand	1:20
7	15	kneeling with support, right hand	1:20
8	7	crouch	0:24

No feedback was available to a student until after all 8 tables (80 trials) had been fired. Visual sighting by the student of bullet holes in the target was not possible. Holes in the targets were covered with black tape by assistants after each score group of 5-shot trials. During this time, the students had their backs to the targets and were reloading for the next score group.

One 5-round magazine was fired for a given table. While the student reloaded, the score (from 0 to 5 hits) was recorded. Another 5 rounds were then fired, the score was recorded, and the next table was fired. Thus, there were two scores from 0 to 5 for each table.

The data were rescored for sequential testing purposes by dichotomizing each 5-round score according to an 80% criterion (0-3=0; 4-5=1). The new set of 48 dichotomous scores for each student was then considered in the sequence in which they were fired, using the sequential testing procedure with the same parameter values described earlier. Thus, a sequential testing-based pass or fail decision was available for each student. For purposes of comparison, each student's score on the total of the 240 shots was dichotomized (0-191=0; 192-240=1) to yield an overall pass or fail decision.

Results. The 48 scores were insufficient to classify 12 students using the sequential procedure. While it is possible to classify each student on the basis of the probability ratio after all the data have been analyzed, such classifications must be considered as a special case since they alter the previously specified values of the necessary parameters. Table 4 compares the classifications of the 225 students for whom the sequential procedure did lead to a decision with the overall test decision.

Table 4
Sequential and Overall Decisions for the MPFQC

	Sequential Decision		
	Pass	Fail	Total
Overall Pass	45	39	84
Decision Fail	1	140	145
Total	46	179	225

False Positives: 1 or 0.40%
False Negatives: 39 or 17.3%
Total Misclassifications: 40 or 17.8%
Average Number of Items to Reach a Decision: 13.77

The results are encouraging from the point of view of the sample size but both disappointing and confusing with respect to misclassification error. An average of 13.77 observations was required for a decision, which is less than one-third of the total number of 48 observations. Misclassification

errors were higher than would be expected from the initial parameters, as well as being in the opposite direction from them. Initially, α, the false negative risk, was set at .01; and β, the false positive risk, was set at .10. The data, however, show a false negative value of .17 and a false positive value of .004. Explanations of these results may lie in violations of the assumptions that (1) the order of item presentation does not introduce bias into the decisions and (2) for any given student, the probability of correctly responding is the same for all items and is equal to that student's hypothetical proportion of correct responses over the entire domain.

Actually, the test could be divided into three subtests. The first eight items in each repetition were similar and relatively difficult, with a mean proportion of hits equal to .64 and a standard deviation of .05. The next six items in each repetition were again similar, but considerably easier, with a mean of .87 and a standard deviation of .04. The last items in each repetition were extremely easy with a mean of .97 and a standard deviation of .002. Thus, each student was faced with a series of difficult items followed by some easy ones, some difficult ones, and so forth. Because of the relatively high passing criterion and the nature of the sequential test, a series of difficult items would tend to favor rejection decisions, particularly if they appeared early in the test--hence, the preponderance of negative decisions observed.

Re-analysis. In an attempt to identify the effect of the inappropriate sequence of items on the sequential test, the data were re-analyzed after partially randomizing the sequence. Reference to a table of random numbers led to the following sequence in terms of MPFQC table numbers: 5, 7, 1, 8, 2, 4, 6, 3. The two scores from each table were again sampled in their original sequence and the above randomized sequence repeated for the three repetitions of the test. The results show some differences.

Although the total number of false negatives decreased, there were increases in the number of students for whom no decision could be made and the average number of observations required. After removing the students for whom no decision was possible, 214 remained. There was again one false positive decision (β observed=.005). Twenty-three false negative decisions (α observed=.107) were observed. The average number of observations increased to 18.39.

The decrease in the number of false negative decisions, along with the increase in the average number of observations and no decisions, tended to favor the conjecture that the original sequence of items led to too many rejection decisions too early. The disappointing results, even after the partial randomization, suggest that the extreme differences in item difficulty severely degrade the applicability of the sequential procedure for data of this type. A clear need for research investigating the robustness of the procedure is indicated.

Performance Testing

Data. The third example examined an application of sequential analysis in the individual Army training and testing context (Knerr & Epstein, 1976). Specifically, the effectiveness of sequential analysis was compared to the total score on the same test as that used for the classification decision and to a criterion external to the test itself. In addition, the efficiency of the sequential testing procedure was measured using the percent of the total number of test items required to make the sequential decisions. The overall objective, then, was to explore the combined effectiveness and efficiency of sequential analysis for the proficiency classification of individuals.

Data for 500 enlisted men at 3 levels of training on 5 combat topics were available from previous research (Knerr, Downey, & Kessler, 1975). In that study, hands-on performance tests were administered approximately one week after the training phase of the research. The topics and number of soldiers tested in each were: Hand Grenades, n=107; Light Antitank Weapon (LAW), n=112; M16A1 Rifle, n=117; Mortar Fire Direction Computer (FDC), n=105; and Surveyed Firing Charts, n=59.

Scores were available for each soldier on each item and on the total test. An index of training was established based on the extent of refresher training that the soldiers received during the research. The lowest level was "no refresher training" (scored 1), the middle level was "some refresher training" (scored 2), and the highest level was "the most effective refresher training" (scored 3). On tests administered immediately after the refresher training, soldiers who received the most effective training scored 82%, on the average, while soldiers who received some training scored 55% correct, on the average.

The sequential analysis procedure was applied to the item data using the parameters described earlier. The items were analyzed in the order that they were administered so that the classification decisions were those that would have resulted if the sequential procedure was applied during the actual test administration. The classification decision for each examinee and the number of items required to reach the decision were noted. In some cases, no decision was made before the total number of test items was exhausted. When no decision was made, the maximum number of test items (total items in the test) was recorded for the examinee. Score codes were 1 for non-proficiency, 2 for no decision, and 3 for proficiency.

Results. The sequential analysis procedure classified 93% of the examinees as either proficient (passed) or non-proficient (failed). No classification decision was made for 7% of the examinees. The Mortar FDC test produced sequential decisions in all cases, and the M16A1 Rifle test produced decisions for all but one examinee. There was no decision reached for between 13-14% of the examinees on the other three performance tests.

Over all 500 examinees, 31% passed and 62% failed according to the sequential decision procedure. It is important to note that the tests on

which the sequential decisions were based were administered approximately one week after the training, so that the scores the trained examinees earned were lower than the typical 80% correct scores. In contrast, the scores immediately after training gave an index of the effectiveness of the training and were closer to the typical post-training level. Generally, soldiers who received the most effective training more often passed; and soldiers who received no refresher training more often failed.

The correlations between the sequential decisions and the training index (Table 5) ranged from .20 (LAW test; $p<.05$) to .47 (Hand Grenade test; $p<.01$); the average of the correlations was .34. The correlations between the total test scores and the training index ranged from .27 (M16A1 Rifle test; $p<.01$) to .58 (Hand Grenade test; $p<.01$), with an average correlation of .38. Thus, the sequential decisions were about as effective as the total test scores when effectiveness was assessed against a criterion external to the text.

Table 5
Intercorrelations, Means, and Standard Deviations of Sequential Decisions, Total Test Scores, and Training Index

	Correlation with Training Index		Means and Standard Deviations					
			Sequential Decision		Total Test Score		Training Index	
Test Topic	Sequential Decision	Total Test Score	Mean	*SD*	Mean	*SD*	Mean	*SD*
Hand Grenades	.47**	.58**	1.26	0.57	15.04	6.10	2.06	0.82
LAW	.20*	.34**	1.55	0.83	14.44	4.86	2.02	0.83
M16A1 Rifle	.25**	.27**	1.69	0.95	41.85	11.75	1.98	0.83
Mortar FDC	.40**	.39**	1.88	1.00	21.48	21.73	2.00	0.82
Surveyed Firing Charts	.35**	.28*	2.46	0.82	25.44	8.55	1.95	0.81

**$p<.01$
*$p<.05$

Since decision accuracy (the extent of classification error) is known to be a problem when the classifications are based on a small number of items, the agreement between the sequential decisions and the total test scores was examined. Where the sequential analysis produced unambiguous proficiency decisions, these decisions agreed with classifications based on the total test score (using the 80% correct criterion) for 88% of the examinees. Thus, the classifications were in error for 12% of the examinees.

Examining just the erroneous decisions, 11% were false positives and 1% were false negatives. These error rates corresponded closely to the predetermined allowable error rates, α and β. The value of α was set at .01, and the extent of false negative decision error was restricted to this amount in these data; β, set at .10, was closely approximated by the obtained error rate for false positives, 11%. Thus, in empirical data on a sample of 500 soldiers, the observed error rates were close to the predetermined ones, indicating that the sequential procedure functioned properly with regard to control over the degree of classification error.

The efficiency of the sequential procedure was examined by comparing the total number of items in each test with the percent of the total number of items required to reach the sequential decisions (Table 6). Averaged over individual examinees, between 19% (Mortar FDC test) and 66% (LAW test) of the total number of test items were required for the sequential decisions. Over the five performance tests, an average of 33% of the items were required to make the sequential decisions. Thus, by using the sequential analysis procedure rather than the traditional procedure of administering the entire test, two-thirds of the items, test administration time, and cost could have been saved. It appears that by combining the accuracy and efficiency results, little accuracy is gained by administering the total test rather than the smaller portion required for the sequential decisions.

Table 6
Items Required for Sequential Decisions and Test Internal Consistency

	Test Items Required for Sequential Decision		
Test Topic	Average Number	Percent of Total	Internal Consistency (KR20)
Hand Grenades	12.51	43%	.86
LAW	16.41	66%	.83
Mortar FDC	10.43	19%	.99
M16A1 Rifle	15.68	24%	.93
Surveyed Firing Charts	16.02	49%	.96

These data also demonstrated the influence of test homogeneity, or internal consistency, on the sequential decision outcomes. The mathematical basis of the sequential analysis method assumes that the observations are randomly sampled from a single domain. In the present application, the items in a test should all measure the same skill. The order of item administration should be irrelevant, since for any item, passing or failing may indicate the classification decision for the total test. That is, the pass or fail score on any item should correspond to the pass or fail classification decision.

This correspondence implies internal consistency, as measured by the Kuder and Richardson (1937) Formula 20, for which the results are reported in Table 6. The Mortar FDC test had the highest internal consistency (.99), and in the Mortar FDC data the sequential analysis procedure functioned very well. It classified all of the soldiers using only 19% of the test items to reach the decisions, and the sequential decisions were highly related to the training index. Thus, the sequential procedure was accurate and efficient in data that met the homogeneity assumption.

In contrast, the LAW test had lower internal consistency (.83), and in the LAW data the sequential procedure did not work as well. More items

were required to reach the decisions (66%), a high number of soldiers had no sequential decision (13%), and the decisions did not relate highly with the training index. In short, when the items did not measure the same underlying skill, or did not measure it reliably, a higher portion of the items was required to reach the decisions; and for a higher portion of examinees, no decision was reached.

The sequential analysis procedure was more effective for tests that measured a single objective and comparatively less effective for tests that either measured diverse objectives or were less reliable. If the scores have low internal consistency for any reason, the order of presentation of the items may have an effect on the decisions. This violates the sequential testing assumption that the observations are a random sample drawn from the same domain.

Conclusions

Wald's sequential probability ratio test for testing the mean of a binomial distribution appears to offer a method for making educational decisions. If the testing situation can be logically interpreted as testing the hypothesis that student performance or instructional effectiveness does not exceed some criterion probability of failure, then the sequential probability ratio test may prove useful. The procedure forces decision makers to explicitly define their definition of acceptable instruction or performance and to consider the misclassification risks that are always present in incomplete sampling. Bringing such issues into the open will certainly be of benefit to the testing community.

The examples included in this paper and other published work (Kriewall, 1969; Linn, Rock, & Cleary, 1972) indicate that the procedure is efficient in terms of the number of observations required for decisions. When the assumptions of the model are met, the procedure appears to be reasonably valid. The major theoretical difficulty lies in insufficient knowledge of the procedure's robustness to violations of the assumptions. On a practical level, the procedure should be easy to implement as part of a computer-managed tested program. It may also prove to be useful (1) in testing situations requiring that students demonstrate their skills at individual stations as part of a performance test; (2) in evaluating instructional effectiveness using small group tryouts; or (3) in other cases where it is practical to consider a decision after each (or perhaps after each small group) of observations. What is now needed, rather than re-analyzing existing data, is experience in using the procedure directly for decision making to determine its proper place in the collection of methods available to decision makers.

References

Baker, E. L., & Alkin, M. C. Annual review paper: Formative evaluation of instructional development. Audio-Visual Communication Review, 1973, 21, 389-418.

Epstein, K. I. Sequential plans and formative evaluation. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, April 1975.

Knerr, C. S., Downey, R. G., & Kessler, J. J. Training individuals in Army units: Comparative effectiveness of selected TEC lessons and conventional methods (ARI Research Report 1188). Washington, DC: Army Research Institute for the Behavioral Sciences, December 1975.

Knerr, C. S., & Epstein, K. I. Sequential analysis for individual proficiency decisions. Paper presented at the annual meeting of the Military Testing Association, Gulf Shores, AL, October 1976.

Kriewall, T. Applications of information theory and acceptance sampling principles to the management of mathematics instruction (Technical Report 103). Madison, WI: The Wisconsin Research and Development Center for Cognitive Learning, October 1969.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test rehability. Psychometrika, 1937, 2, 95-101.

Linn, R. L., Rock, D. A., & Cleary, T. A. Sequential testing for dichotomous decisions. Educational and Psychological Measurement, 1972, 32, 85-95.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement (American College Testing Technical Bulletin 18). Iowa City, IA: American College Testing Program, January 1974. [Also in C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement (CSE Monograph Series in Evaluation 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.]

Wald, A. Sequential analysis (2nd ed.). New York: Dover Publications, Inc., 1973. (Originally published, 1947)