The Number of Performance Assessments Necessary to Determine Competence

Robert L. Lathrop
Professor, Vocational Education
Florida State University
Tallahassee, FL 32306

Abstract. Previous approaches to the issues of misclassifications and test length in criterion (competency, domain)-referenced assessment have usually been based on binomial or on Bayesian probability models which are not directly applicable to most instruction/selection decisions. The logic of the Sequential Probability Ratio Test, however, exactly parallels competency or mastery decision problems. It does not require a priori knowledge of individual examinees nor information about prior performance of other groups. It allows the examiner to establish reasonable levels of acceptance and rejection errors and specifies the minimum number of testing trials necessary to make individual decisions within these limits. The procedure, although mathematically unfamiliar to many practitioners, lends itself to a graphic solution with the use of only two tables.

Background

Every day each of us makes countless decisions about others based on our observations of their behavior. For those of us in educational enterprises, many of these decisions are judgments about another person's level of competence. In recent years educators have been interested in an ordinal measurement scale which describes individuals as having, or not having, competence (mastery) of some particular skill or knowledge.

The underlying problem faced by an educator making competency decisions is to attempt to describe each examinee's *true* level of competence, that is, to correctly describe the individual as *competent* or *not competent* according to a predetermined criterion. Because we are dealing with fallible assessment procedures and human inconsistency, some degree of decision error or misclassification is inevitable. The issue addressed in this presentation is basically a sampling question: How many observations of an individual's behavior are necessary to provide a prespecified risk of an erroneous decision about competence?

Although a number of writers have discussed this subject, their approaches have generally not proved useful to practitioners (Millman, 1972, 1973; Novick and Lewis, 1974; Hambleton, et al., 1976, 1978). Rather than describe these approaches to determining test length (number of assessments), we will instead consider an alternative approach developed by Wald (1947) as a means of making quality control decisions. Wald's approach, known broadly as sequential analysis, seems to fit competence assessment problems perfectly.

Consider a situation where an instructor wishes to make mastery decisions about individual learners. We will make the assumption that at any given point in time a particular learner either *has* mastery or *does not have* mastery, according to some predetermined criterion. The instructor's task is to make decisions about individual learners in such a way that over a large number of such decisions, (s)he makes no more than some preagreed upon number of misclassifications. The reader should bear in mind that for each individual we make only one decision (at a time) and therefore, we are either right or wrong in that particular choice. What we would like is a decision rule that allows us to control the risk of making incorrect decisions.

Consider Figure 1, which illustrates the possible outcomes for mastery decisions. On the left margin we assume that each individual either has mastery or

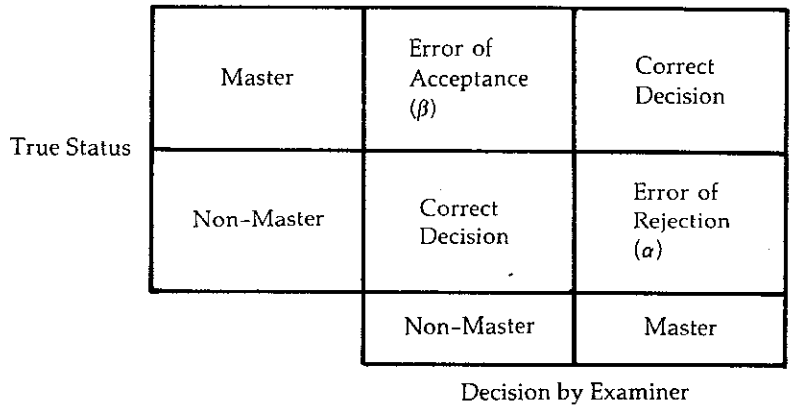

Figure 1. Possible outcomes of a binomial decision.

does not have mastery according to our definition. Obviously we do not know for any individual which category is correct for him/her. We do, however, have the results from our testing, and it is on this basis that we make the decision shown on the baseline of Figure 1.

If the examinee has *Mastery* and we so classify him/her, we have made a correct decision. If, on the other hand, our assessment leads to the conclusion that (s)he does not have mastery, we have made an error of "rejection." In hypothesis testing this type of error is referred to as Type I or Alpha Error.

Now consider the converse. Suppose that the individual *does not have Mastery*. If we conclude, based on our assessment, that the individual does not have mastery, we have made a correct decision. It is possible, however, that our assessment erroneously leads us to conclude that the individual does have mastery. In this case, we would have made an "acceptance" error. This type of error is customarily referred to as a Type II or Beta Error.

Although we speak of Type I and Type II errors (or misclassifications), such errors exist only in an inferential sense. We do not know the "true" mastery status of an individual. Comparisons of our decisions to "true" status, therefore, are always inferences. To the extent that we act on the basis of our decisions as if they were true, the "true" status of an individual may be academic. In a baseball game a called pitch *is* a "ball" or a "strike" depending on what the plate umpire says it is. If this is the mode of using assessment data to make instructional decisions, "true" status may be conceptually interesting, but of little operational value. As the expression goes, "They is what I says they is!"

In a broader sense, however, erroneous decisions always have consequences whether we acknowledge them or not. In a recreational game a bad call may affect the final score. In an employment decision a "bad call" may result in employing a noncompetent worker (or rejecting a competent one). In an instructional decision a misclassification may require a learner to remain at some level of instruction longer than necessary (or, conversely, allow a noncompetent learner to advance prematurely). In the final analysis, all misclassifications have consequences. Whether a particular misclassification is important is determined by the seriousness of the consequences. In some instances, the consequences of a classification error are borne primarily by the individual being assessed and may be inconsequential to the institution making the error. In other cases, both the individual and the institution (organization, agency, etc.) making the decision bear the costs of erroneous decisions. There are always costs for misclassifications whether we are conscious of them or not.

On the assumption, then, that we wish to exercise some conscious control over misclassifications, we would prefer a decision strategy that allows us to assign error risks in advance and make our decisions within these limits. The Sequential Probability Ratio Test developed by Wald (1947) meets this requirement. This procedure makes no a priori assumption about the level of performance of an individual examinee, but rather depends on the principle that an individual's level will emerge from a relatively small number of trials. As decision makers, we want a strategy that requires us to draw the minimum number of samples of behavior to categorize an individual's performance within whatever risk limits we have determined are reasonable. Stated another way, a good decision making strategy will allow us to quickly identify an individual who is clearly above or below some prescribed performance standard. The Sequential Probability Ratio Test (SPRT) establishes boundary values based on preselected levels of alpha and beta. The assessment of the individual proceeds (sequentially) until the pattern of assessments crosses either the acceptance or rejection boundary values.

Although it has sometimes been mistakenly believed that the Sequential Probability Ratio Test requires the assumption of independent observations, such is not the case. Wald points out that the basic inequalities on which the test is based are equally valid for dependent observations so long as the probability is 1.0 that the procedure will eventually terminate (pp. 43-44). The fact that the procedure is valid for a very general class of situations allows us to use the technique in practical testing situations where assumptions of independence of observations would not be reasonable.

Procedure

Suppose we wish to determine whether or not an individual learner has reached criterion performance on a particular learning task. We decide that to be competent (have mastery, reach criterion,...) (s)he must attain 80 percent of the maximum possible score on a performance checklist. We also determine, based on our subjective analysis of the consequences, that we wish to hold the risk of allowing a student without mastery to proceed to the next task, to a probability of .10 (error of acceptance, beta = .10). Similarly, we decide we want to set the risk of mistakenly holding back a student who has mastery to a probability of one-in-five (error of rejection, alpha = .20). Having made these determinations we can proceed.

Although the (SPRT) can be presented in either tabular or graphic forms, the writer recommends the graphic approach for the non-mathematically oriented user.

The graphic solution involves plotting two straight boundary lines on an X-Y graph. The horizontal axis is labeled m or (trials) and is simply the number of sequential observations (assessments, testings, trials...) for a particular examinee. This number will vary from individual to individual. The vertical axis is labeled dm, and is the cumulative number of nonpassing trials for an individual examinee (See Figure 2).

The two sloping lines, labeled L_1 and L_0, represent the two boundary lines for our decisions. If we plot m and dm for

any individual examinee, the plotted points will always begin between the base line, and L_1. As the testing proceeds the plotted points for each examinee will eventually cross either L_1 or L_0. As soon as the plotted points cross either boundary line, the testing is complete and an appropriate decision made.

Because L_1 and L_0 are straight parallel lines we need to calculate only three values: the point where L_1 crosses the vertical axis, the point where L_0 crosses the vertical axis, and the slope of L_1 and L_0. Since L_1 is the *rejection* (nonmastery) boundary line, we will label this intercept r_1. Similarly, the point where L_0 crosses the vertical axis is the *acceptance* (mastery) intercept which we will label a_0. Since the lines are parallel, they will both have the same slope, labeled s. To establish line L_1 on our graph, we need to plot two points. We will plot the intercept (where m=0) and the value for m=10.

The mathematical expressions for r_1 a_0 and s are shown by Wald to be as follows:

$$r_1 = \frac{\text{Log}_{10}\left(\frac{1-\beta}{\alpha}\right)}{\text{Log}_{10}\left(\frac{P_1\,[1-P_0]}{P_0\,[1-P_1]}\right)}$$

$$a_0 = \frac{\text{Log}_{10}\left(\frac{\beta}{1-\alpha}\right)}{\text{Log}_{10}\left(\frac{P_1\,[1-P_0]}{P_0\,[1-P_1]}\right)}$$

$$s = \frac{\text{Log}_{10}\left(\frac{[1-P_0]}{[1-P_1]}\right)}{\text{Log}_{10}\left(\frac{P_1\,[1-P_0]}{P_0\,[1-P_1]}\right)}$$

Where:

α = rejection error level
β = acceptance error level
$1-P_0$ = upper limit of criterion tolerance (indecision) region
$1-P_1$ = lower limit of criterion

Although these expressions may appear somewhat intimidating, fortunately they lend themselves to rather simple solutions which can be tabled for commonly used values.

Before proceeding with our example, the preceding mathematical expressions contain two terms, P_0 and P_1 which deserve further comment.

The context of Wald's work was the need to make sampling decisions about the ratio of defective to nondefective parts in industrial shipments. Wald suggested that in most practical situations the decision maker was less interested in knowing the "true" ratio of defects in a shipment than (s)he was setting some upper and/or lower boundary for the "true" percentage. That is, that P, the "true" percentage of defects is no greater than P_1 nor less than P_0. As long as $P_0 < P < P_1$ Wald stated that the decision maker was *indifferent* to the "true" value of P. Within this "zone of indifference," to use Wald's term, there is no practical difference between P and P_0 or P_1.

Obviously the choices of values for P_0 and P_1 are judgmental matters which go beyond pure statistical considerations. In the context of performance testing, if we set mastery at 80 percent, P becomes 20 percent. That is, we will allow 20 percent "defective" performances and still accept the examinee as competent (having mastery). The examiner then has a choice of how closely (s)he wishes to hold to exactly 80 percent. Is the distinction between 79 percent and 80 percent meaningful (or, in terms of P and P_1, is the difference between 21 percent and 20 percent significant)? In a more general sense, how far can P_0 and P_1 range away from P before the difference has practical significance? In most educational testing the necessity to make extremely precise estimates of the "true" value of P for an individual is less important than the ability to quickly sort the examinees into three broad categories:

1. Those individuals who clearly are well above the minimum requirements for mastery;

2. Those individuals who are well below the minimum requirement for mastery;

3. Those individuals who require additional testing before a mastery determination can be made.

The more precision we require (that is, the closer we set P_0 and P_1 to P), the larger must be the number of testing trials. In the examples and tables developed for this presentation, the writer has set P_0 = P - .10 and P_1 = P + .10. Translated into the context of this example if:

P = .20
1 - P = .80 Nominal criterion level for mastery
P_0 = .10
1 - P_0 = .90 Upper limit of Region of Indecision (Zone of Indifference)
P_1 = .30
1 - P_1 = .70 Lower limit of Region of Indecision (Zone of Indifference)

By setting .10 < P < .30, we have, in effect, said we want to quickly identify examinees who can perform acceptably on 90% of the trials (with a misclassification risk of beta). Similarly, we want to quickly screen out examinees who cannot perform acceptably on as many as 70% of the trials (with an accompanying misclassification risk = alpha). The remaining group whose performance is above 70% but less than 90% we say is "too close to call" without further testing. The practical question the examiner must then ask and answer is, "Does the decision justify additional testing or am I willing to make a decision based on the information available, taking into account the seriousness of the consequences of misclassification?" In most performance testing problems familiar to the writer, the values proposed for P_0 and P_1 are a reasonable compromise. The Sequential Probability Ratio Test procedure allows the user to set whatever values for P_0 and P_1 (s)he chooses. The tables presented here, however, apply only to P_0 = P - .10; and P_1 = P + .10.

Table 1 presents intercept values for r_1, and a_0 for several combinations of alpha and beta. The values in Table 1 are based on a criterion level of 1 - P = .80. Because decisions for individuals having mastery levels slightly above or slightly below .80 will be very difficult to determine, we will establish a set of tolerance limits around 1 - P within which we say decisions are not practically possible. In Table 1 the tolerance limits have been chosen as (1 - P) ± .1, making 1 - P = .7; and 1 - P_0 = .9. From Table 1 we locate intercept values r_1 and a_0 for alpha = .20; beta = .10.

r_1 = 1.11
a_0 = -1.54

We also find from Table 1 that

s = .186.

The values for r_1 and a_0 are plotted on the vertical axis of Figure 2 where m = 0. Because of the ease of calculation we will calculate the values of L_1 and L_0 when m = 10.

$L_1 = r_1 + m(s) = 1.11 + 10(.186) = 2.97$

$L_0 = a_0 + m(s) = -1.54 + 10(.186) = .32$

Plotting these values we have the lines L_1 and L_0 shown in Figure 2.

We are now ready to begin our sequential testing. We administer our assessment procedure to Student "A." One possible outcome of this assessment is "pass," if the student meets or exceeds the 80 percent criterion. In this case his/her dm score is 0. If the student does not reach criterion on this trial his/her dm score = 1. No decision about mastery is possible on the first trial since both dm = 0, and dm = 1, fall between L_0 and L_1.

We proceed to trial number two (m = 2). The possible cumulative outcomes are:

dm = 0, if the student reaches criterion on both trials

dm = 1, if the student reaches criterion on one trial

dm = 2, if the student reaches criterion on neither trial

Since dm = 2 for trial two falls above L_1, we would reach a decision of *nonmastery* for any student with a dm score of 2 on the second trial. Students with scores of dm = 0, or dm = 1 on the second trial would continue to be tested.

The sequential testing continues either until the plotting of dm scores for an examinee falls outside L_1 or L_0 or the examiner decides to terminate the testing because of practical limitations. It can be shown, theoretically, that every dm plotting will eventually cross either L_1 or L_0. In practice, however, this number of trials may not be justified. The failure of a dm plot to quickly move out of the region of indecision and cross L_1 or L_0 indicates either that the examinee's true level of performance is very close to the criterion level, or that the assessment procedure is somewhat unstable (unreliable).

Further examination of Figure 2 reveals that at this criterion-level (1 - P = .80), decisions involving mastery (crossing L_0) require more trials, than decisions of nonmastery (crossing L_1). Note that, not until an examinee has performed at or above criterion levels for nine successive trials would his/her dm plot cross line L_0. On the other hand, for an individual with as few as two unsuccessful trials, we could make a nonmastery decision.

From Table 1 we can see that if we require alpha or beta to be small (low risk of a misclassification), the intercepts and therefore lines L_1 and L_0 are further apart. The effect of decreasing the acceptable error limits, therefore, is to increase the number of trials (assessments) needed to make a decision. Conversely,

Table 1
Slope and Intercept Values r_1 and a_0 for Various Levels of α and β. Where $1 - P_0 = .9$; $1 - P_1 = .7$

β Level	a_0 / r_1	α Level .1	.2	.3	.4	.5
.1	r_1	1.63	1.11	.814	.601	.435
	a_0	−1.63	−1.54	−1.44	−1.33	−1.19
.2	r_1	1.54	1.03	.727	.513	.348
	a_0	−1.11	−1.03	−9.28	−.814	−.679
.3	r_1	1.44	.928	.628	.415	.249
	a_0	−.814	−.727	−.628	−.513	−.378
.4	r_1	1.33	.814	.513	.300	.135
	a_0	−.601	−.513	−.415	−.300	−.165
.5	r_1	1.19	.679	.378	.165	.000
	a_0	−.435	−.348	−.249	−.135	.000

Slope = .186

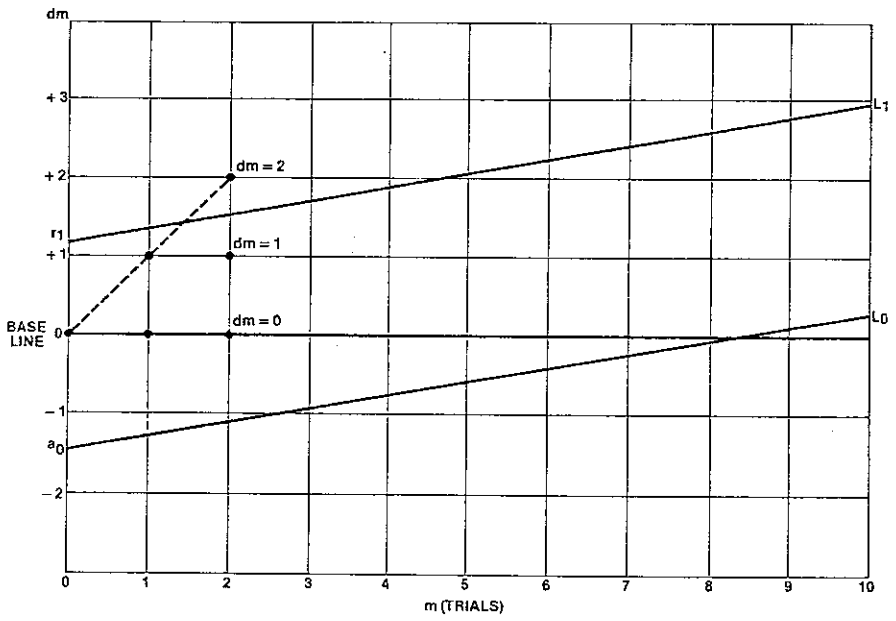

Figure 2. Graphic solution to the sequential probability ratio test for $\alpha = .2$; $\beta = .1$; $1 - P_0 = .9$; $1 - P_1 = .7$

the number of trials needed to reach a decision decreases as we tolerate larger risks of one or both types of misclassification. Because the Sequential Probability Ratio Test considers alpha and beta errors individually, we have a choice of which combination of alpha and beta fits our particular decision context. If, for example, we are principally concerned with not labeling a person with mastery as a "nonmaster," we would set alpha as low as practical. We might in this case allow beta (labeling a nonmaster as a "master") to go to .5, the level of chance. Plotting the intercepts for alpha = .1 and beta = .5 on Figure 3, we see that with as few as three trials, a *mastery* decision could be made (solid lines L_1 and L_0).

If, in this example, we had been primarily concerned with not labeling a nonmaster as a "master" and were

relatively indifferent to overlooking some individuals with mastery (say, in a selection situation where we have more qualified applicants than positions), we could set beta = .1 and allow alpha to operate at a chance level (alpha = .5). In this case, plotting L_1 and L_0 (dotted lines) for beta = .1 and alpha = .5, we find that we can reach some *nonmastery* decisions on the first trial, but would not make any *mastery* decisions with less than seven successive passing trials (dm = 0; m = 7).

In each of the above examples we have used a criterion standard of .80 with a tolerance region of plus or minus .10. Changing the criterion standard for mastery from .80 to another value will affect both the slope and the intercept of L_1 and L_0. Although separate tables parallel to Table 1 could be constructed for each possible criterion-level, the relationship between criterion level and slope and criterion level and intercept can easily be shown in a second table (See Table 2).

Consider an example where the criterion for mastery was .70 instead of .80. To compute intercepts r_1 and a_0, we extract the appropriate values from Table 1 for our selected levels of alpha and beta and multiply these values from Table 1 by the intercept weight shown in Table 2. For example, if we set alpha = .20 and beta = .30, the values from Table 1 are a_0 = -.727 and r_1 = .928. If we multiply each of these values by 1.38 we have a_0 = -1.00 and r_1 = 1.28. The slope is read directly in Table 2 as, s = .293.

We can now plot L_1 and L_0 as shown in Figure 4. For comparative purposes L_1 and L_0, corresponding to a criterion level of .80, will be shown as dotted lines in Figure 4.

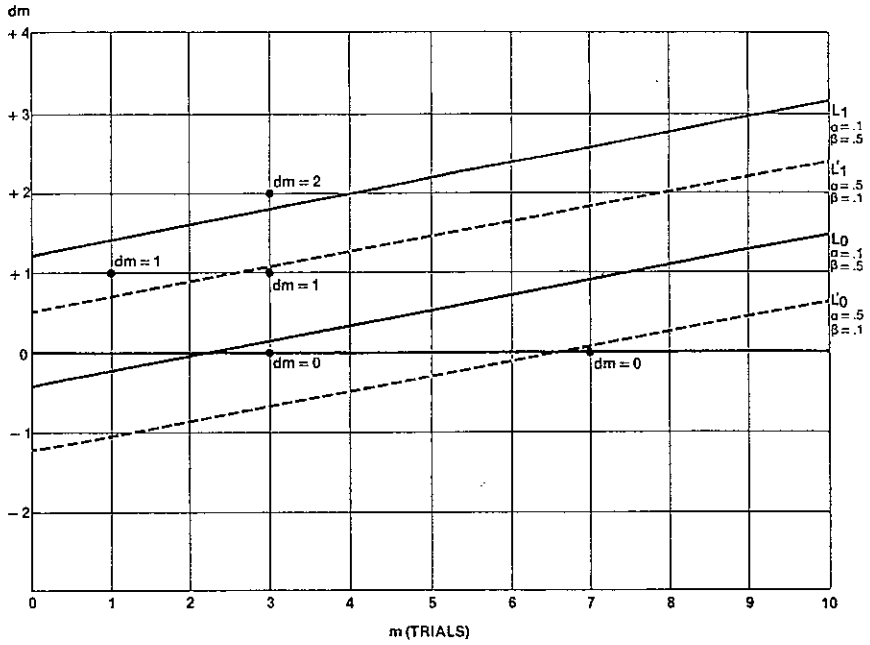

Figure 3. Graphic solution to the sequential probability ratio test comparing $\alpha = .1$; $\beta = .5$; with $\alpha = .5$; $\beta = .1$

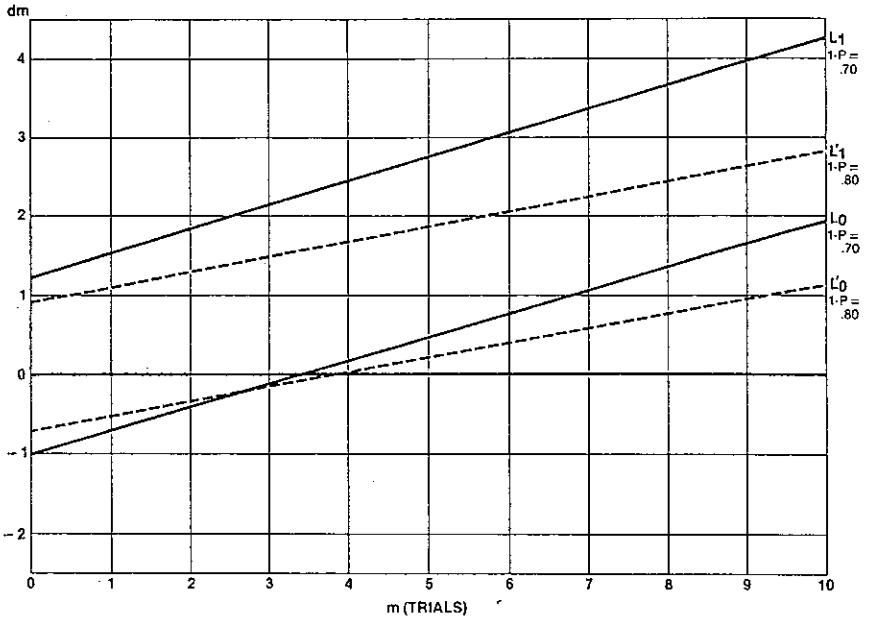

Figure 4. Graphic solution to the sequential probability ratio test comparing $1 - P = .70$ with $1 - P = .80$; $\alpha = .2$; $\beta = .3$; $1 - P_0 = .9$; $1 - P_1 = .7$

Summary

Previous approaches to the issues of misclassifications and test length in criterion (competency, domain)-referenced assessment have usually been based on binomial or on Bayesian probability models which are not directly applicable to most instruction/selection decisions.

The logic of the Sequential Probability Ratio Test parallels, exactly, the logic of most competency or mastery decision problems. It does not require a priori knowledge of individual examinees nor information about prior performance of other groups. It allows the examiner to establish reasonable levels of acceptance and rejection errors and specifies the minimum number of testing trials to make an individual decision within these limits. The procedure, although mathematically unfamiliar to many practitioners lends itself to a graphic solution with the use of only two tables. Once plotted, the selection consequences of various choices for alpha, beta, criterion level (1 - P), and tolerance limits become obvious.

The practitioner may find several of the following conclusions helpful in his/her use of the Sequential Probability Ratio Test to make instructional/placement/selection decisions regarding mastery.

1. The choice of levels for alpha and beta should be based on realistic estimates of the consequences of a particular misclassification. In some selection problems the "cost" of an error may fall primarily on the institution doing the selection as in the case where an applicant who is not qualified is selected. In others the "cost" may be borne entirely by the qualified examinee who is not selected. Both individual and institutional costs should be considered in setting alpha and beta levels.

2. In all cases, alpha and beta should be allowed to be as large as practical, if

Table 2

Slope and Intercept Weighting Factors for Several Criterion Levels*

Criterion Level	Slope(s)	Intercept Weight
.50	.500	1.66
.60	.397	1.59
.65	.346	1.50
.70	.293	1.38
.75	.241	1.21
.80	.186	1.00
.85	.128	.73

*Based on tolerance factors of $\pm .10$.

decisions are to be made in the minimum number of trials. If alpha and beta must be kept at very low levels, the number of trials to make such decisions must be expected to increase accordingly.

3. In many testing problems we are interested in controlling only one type of error. If, for example, we are screening individuals on a pretest, we may be relatively indifferent to making a Type I error (not exempting a student who already has mastery) on the argument that even the student who has minimal mastery could benefit from some "overlearning." In selection decisions examiners are typically interested in minimizing Type II errors (erroneously selecting a nonqualified employee, admitting a nonqualified student, etc.). Where classification decisions can be revised based on subsequent observations, the consequences of either a Type I or a Type II error can be minimized and therefore initially allowed to remain somewhat large.

4. Having established levels for alpha and beta, the number of testing trials becomes a function of the tolerance limits for the region of indecision, $(1 - P_0)$ and $(1 - P_1)$. If the nature of our decision requires us to make discriminations close to the specified criterion level, that is, the values $(1 - P_0)$ and $(1 - P_1)$ are set numerically close to the criterion level $(1 - P)$, then we must expect to collect a large sample of test behavior before reaching a decision. Conversely, if we are willing to classify individuals somewhat broadly and are willing to suspend decisions for persons with test results close to the criterion, we can reduce the number of testing trials substantially. In all cases we should allow the range of $(1 - P_0) > (1 - P) > (1 - P_1)$ to be as wide as possible.

5. In virtually all cases, more than one trial or performance assessment is required to reach a decision about mastery. Under certain circumstances a decision that an individual does *not* have mastery can be reached in only a few trials, perhaps, two or three. In contrast, rarely would we be able to decide that an individual had mastery with less than four or five successful trials.

6. There is no way in advance to determine how many trials will be required to make a decision about an individual examinee. The process, however, can be discontinued at the end of some predetermined number of trials, with those examinees for whom a clear pattern of performances has not emerged being placed with whichever decision has the less serious consequences.

7. The number of trials necessary to make a decision is influenced by the choice of the criterion-level. In general, the effect of lowering the criterion-level is to increase the number of trials needed to make a decision. A further consequence of lowering the criterion-level is to increase the slope of the decision limit lines L_1 and L_0. As the slope of the decision limits lines increases the number of trials required for nonmastery decisions *increases* and the number for mastery decisions *decreases*. The reasonableness of this consequence is intuitively obvious; as we lower the level of performance required for mastery we reject fewer persons for lack of competence and accept more persons as having mastery.

In all cases the Sequential Probability Ratio Test provides the smallest number of sample observations necessary to make binomial decisions within predefined error limits. In many cases reliable decisions can be made with half the number of observations suggested by other sampling procedures (Wald, 1947). Because a smaller number of undecided cases remains after each successive testing, the Sequential Probability Ratio Test allows us to quickly and efficiently separate the individuals with consistently high or consistently low performance from those individuals whose performance pattern is not clear-cut. For a given expenditure of testing effort, Cronbach and Gleser (1957) have shown sequential sampling as being the most efficient of several decision strategies compared. The procedure outlined in this presentation will allow this powerful decision-making technique to be used by educational practitioners for a variety of selection/placement decisions.

References

Cronbach, L., & Gleser, G.C. *Psychological tests and personnel decisions*. Urbana: University of Illinois Press, 1957.

Hambleton, R.K., Hariharan, S., Algina, J., & Coulson, D.B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 1978, 48,* (1), 1-47.

Millman, J. *Determining test length*. Los Angeles: Instructional Objectives Exchange, 1972.

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research, 1973, 43,* 205-216.

Novick, M.R., & Lewis, C. Prescribing test length for for criterion-referenced measurement. In C.W. Harris, M.C. Alkin & W.J. Popham (Eds.) *Problems in criterion-referenced measurement*. CSE Monograph Series in Evaluation No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Wald, A. *Sequential analysis*. New York: John Wiley and Sons, 1947.