Sequential Testing for Selection

R. A. Weitzman
Naval Postgraduate School

In sequential testing for selection, an applicant for school or work responds via a computer terminal to one item at a time until an acceptance or rejection decision can be made with a preset probability of error. The test statistic, as a function of item difficulties for standardization subgroups scoring within successive quantiles of the criterion, is an approximation of a Waldian probability ratio that should improve as the number of quantiles increases. Monte carlo simulation of 1,000 first-year college students under 96 different testing conditions indicated that a quantile number as low as four could yield observed error rates that are close to their nominal values with mean test lengths between 5 and 47. Application to real data, for which interpolative estimation of the quantile item difficulties was necessary, produced, with quantile numbers of four and five, even more accurate observed error rates than the monte carlo studies did. Truncation at 70 items narrowed the range of mean test lengths for the real data to between 5 and 19. Important for use in selection, the critical values of the test statistics are functions not only of the nominal error rates but also, alternatively, of the selection ratio, the base-rate success probability, and the success probability among selectees, which a test user is free to choose.

A serious problem of conventional testing for selection is that estimation accuracy is not uniform over the entire range of applicant ability. Conventional tests ordinarily estimate middle-level abilities with less error than abilities at the lower or upper levels. This is particularly a problem if a number of users of a single test wish to select different fractions of applicants, some of which differ substantially from one-half. The percentage of selection errors in the case of any of the more extreme selection fractions may tend to be unacceptably large. Sequential testing, which consists of the presentation via a computer terminal of one item at a time until a function of the item responses reaches a predetermined critical value, has the potential of working to resolve not only this but also another problem that is important in a selection context: the lack of precise control in existing testing procedures over both the probability of accepting an applicant who will fail and the probability of rejecting an applicant who would succeed if accepted.

One form of sequential testing—"adaptive" testing, introduced by Lord (1968) as "tailored" testing—estimates the ability of an applicant by presenting one item at a time so that after the first item the choice of items presented depends on (is tailored to) the applicant's responses to the preceding

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 3, Summer 1982, pp. 337–351

0146-6216/82/030337-1581.75

items. Because each applicant thus tends to respond not only to a different number but also to a unique set of items, adaptive testing can obtain uniformly accurate estimates over the entire applicant ability range. Although an adaptive test used for selection may thus be equally accurate (or inaccurate) regardless of the fraction of applicants selected, it sacrifices in this very equality needed accuracy in the neighborhood of the cutting score for unneeded accuracy elsewhere in the ability range. Adaptive tests also generally neither specify nor control the probabilities of selection errors. A sizable literature exists on adaptive testing; two proceedings edited by Weiss (1978, 1980) contain a fair representation of this literature.

Interest in sequential testing for classifying individuals is not new. Application to dichotomous classification by Linn, Rock, and Cleary (1972) of a sequential procedure developed by Armitage (1950) for polychotomous classification tended to confirm in real-data simulation a theoretical prediction made by Green (1970) that sequential procedures might produce a reduction in testing time over conventional procedures of about 50%. The procedure investigated by Linn et al. (1972) required an end to testing as soon as a statistic computed from the response data was farther from zero than a preset number. Though a monotonic function of this number, the observed rate of classification errors was not subject to precise control by the procedure.

Developed specifically for use in selection, the form of sequential testing described here can both concentrate its accuracy at the cutting score and control the probabilities of selection errors. Called *selective testing,* this form of sequential testing is an adaptation to selection of the sequential probability ratio test (SPRT) developed by Wald (1945). Other testing adaptations of the SPRT apply specifically to the determination of subject matter mastery (Epstein & Knerr, 1978; Ferguson, 1970; Kalisch, 1980; Kingsbury & Weiss, 1980; Reckase, 1980). These adaptations all involve Wald's binominal test of two proportions, typically of the amount of subject matter known, that bracket an "indifference region"; only for students whose proportions fall outside this region do the mastery or nonmastery decisions tend to have error rates that are no higher than previously specified values. Selective testing, by contrast, works to control the error rates for everyone.

Testing Procedure

Most Waldian sequential methods use a probability ratio test statistic updated after each observation to test two *point* hypotheses, such as H_1: $\mu = \mu_1$ against H_2: $\mu = \mu_2$, where μ designates either a population mean or, subscripted, a specific (point) value of the mean. Wald (1945) suggested but did not develop tests of corresponding *composite* hypotheses—H_1: $\mu < \mu_0$ versus H_2: $\mu > \mu_0$. Following directly from Wald's suggestion, which involved integrals in the numerator and the denominator of the test statistic, the procedure proposed here uses summations to approximate the integrals. This procedure is a test for each applicant of the composite hypotheses H_1: $\theta < \theta_0$ versus H_2: $\theta > \theta_0$, where θ designates a measure of the applicant's subsequent school or job performance and θ_0 designates the performance measurement separating success from failure for all accepted applicants. In this procedure, the applicant responds at a computer terminal to one item at a time until, on reaching a value farther from one than an upper or lower critical value, the probability ratio test statistic determines the selection decision with preset probabilities of error: α, the probability of accepting an applicant for whom $\theta < \theta_0$, and β, the probability of rejecting an applicant for whom $\theta > \theta_0$.

The approach to the development of this procedure relies, as large-scale empirical test development generally does, on the existence of both performance measurements and item responses for a large, so-called "standardization" group of individuals. The performance measurements are ordered, and K separate subgroups of individuals having performance measurements in intervals between successive quantiles are identified. If the quantiles are quartiles, for example, $K = 4$, corresponding to

the four subgroups between the zeroth and the first quartile, between the first and the second quartile (median), between the second and the third quartile, and between the third and the fourth quartile. For each subgroup k ($k = 1, 2, \ldots, K$), the proportion of individuals who answer item i correctly is determined: p_{ik}. Then, in subsequent testing, the probability ratio L_n is updated for each applicant as the applicant responds successively to items $i = 1, 2, \ldots, n$:

$$L_n = \frac{(K - K^* + 1)^{-1} \sum_{k=K^*}^{K} \prod_{i=1}^{n} p_{ik}^{x_i} (1 - p_{ik})^{1 - x_i}}{(K^* - 1)^{-1} \sum_{k=1}^{K^*-1} \prod_{i=1}^{n} p_{ik}^{x_i} (1 - p_{ik})^{1 - x_i}}, \qquad [1]$$

where K^* designates the standardization subgroup immediately above the performance measurement (θ_0) separating success from failure and x_i equals 1 for a correct and 0 for an incorrect response to item i. On the local independence assumption that item responses are independent for applicants who have equal performance measurements, this ratio is an approximation that should improve as K increases. To the extent that local independence exists within every subgroup, each product in this ratio is the probability that a member of the corresponding subgroup will make the response sequence $x_1, x_2, \ldots, x_n$. Subgroup memberships are mutually exclusive. Multiplied by $1/K$, which cancels, the numerator of L_n is thus the probability that a member of one of the top $K - K^* + 1$ subgroups will make the response sequence; and the denominator, the probability that a member of one of the bottom $K^* - 1$ subgroups will make the response sequence. According to Wald (1945), the critical values for L_n are $(1 - \beta)/\alpha$ and $\beta/(1 - \alpha)$; acceptance occurs when L_n goes above $(1 - \beta)/\alpha$ and rejection when L_n goes below $\beta/(1 - \alpha)$. These critical values tend to assure the error probabilities, α and β, because approximately $1 - \beta$ of the applicants belonging to the top $K - K^* + 1$ subgroups and α of the applicants belonging to the bottom $K^* - 1$ subgroups will have values of L_n above $(1 - \beta)/\alpha$, while β of the applicants belonging to the top $K - K^* + 1$ subgroups and $1 - \alpha$ of the applicants belonging to the bottom $K^* - 1$ subgroups will have values of L_n below $\beta/(1 - \alpha)$. The qualification "approximately" reflects the tendency of L_n to change discretely, rather than continuously, as a function of n. If K is large enough, therefore, testing ends with error probability approximately α as soon as L_n becomes larger than $(1 - \beta)/\alpha$ or with error probability approximately β as soon as L_n becomes smaller than $\beta/(1 - \alpha)$. As real world estimates rather than theoretical values, however, the p_{ik}'s decrease in reliability as K increases. Fortunately, monte carlo data simulation, described in the next section, indicates that a value of K as low as 4 can yield actual error rates that are close to their nominal values.

Although the items presented in this procedure may vary from applicant to applicant, the test length for each applicant will tend to be shortest if the items presented are the ones for which the differences $p_{iK^*} - p_{i(K^*-1)}$ are largest. Making the item-to-item discrete changes in L_n correspondingly largest, these are the items that are the most discriminating near θ_0. Applicants for whom θ is near θ_0 will generally need to respond to more items than other applicants. Control over test length for all applicants is possible, however, by manipulating the nominal value of β, provided the cost of rejecting applicants who might be successful if selected is not too high. The monte carlo studies to be reported indicate that test lengths between 7 and 11 may provide predictions for which error rates closely approximate the nominal values of $\alpha = .05$ and $\beta = .15$. If the number of acceptable applicants is greater than the required or desired number, the selection procedure can attempt within the acceptable pool to meet goals of racial or ethnic balance. This attempt will be free from charges of reverse discrimination because no test score differences distinguish one acceptable applicant from another.

More familiar than α and β within a selection context are the selection ratio (Ψ); the base rate (Ω), which is the proportion of applicants who would be successful on the criterion if all were selected; and the corresponding proportion for the actual selectees (ω). (In this notation, the lower-case letters are conditional, and the upper-case letters marginal probabilities.) These five probabilities are related so that knowledge of any three of them determines the other two. Because of this relationship, the critical values of L_n are expressible in terms of Ψ, Ω, and ω, rather than in terms of α and β, as follows:

$$A = \frac{\omega(1 - \Omega)}{\Omega(1 - \omega)} \qquad [2]$$

for the upper and

$$B = \frac{(1 - \Omega)(\Omega - \omega\Psi)}{\Omega[(1 - \Omega) - \Psi(1 - \omega)]} \qquad [3]$$

for the lower critical value. Circumstances normally determine the values of Ψ and Ω. Different from a conventional test, in which the predictive validity (together with Ψ and Ω) determines the value of ω, a test user is free to choose this value in a selective test. Affecting only the expected number of items, this freedom of choice is a unique advantage of a selective test, unshared by sequential tests, like the ones cited previously, that do not control overall error rates. Another advantage, of course—one shared by all sequential tests—is the expectation that the number of items corresponding to any value of ω will be lower than for a conventional test.

Monte Carlo Illustration

Method

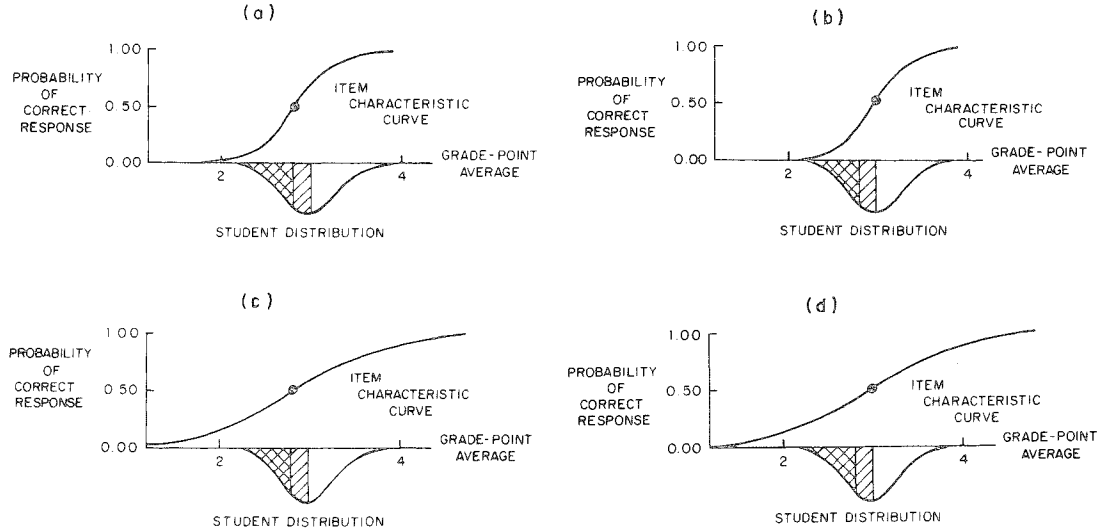
The data simulated consisted of the criterion grade-point averages and predictor item responses of 1,000 first-year college students. Models used to create the data specified the parent distribution of grade-point averages and the theoretical distribution of item difficulties as a function of these averages. Figure 1 shows these distributions. The parent distribution of grade-point averages was normal with a mean of 3 and a standard deviation (σ) of ⅓; the item difficulties followed an ogival item characteristic curve (ICC) over the scale of grade-point averages with a mean at either their 25th or their 50th centile and a standard deviation equal to either σ or 2σ. In all, the 1,000 fictitious college students went through 96 different monte carlo testing simulations: one for each of the four ICCs as well as each of two failure fractions (¼ and ½), also shown in Figure 1; each of two values of K (4 and 12); and each of six combinations of α and β (β = .05 to .15 and α = .05 to β, both in steps of .05).

To compute the probability ratio L_n, p_{ik} was approximated by the weighted average of difficulties for item i within quantile k:

$$\frac{\sum_{j \subset k} \phi(\theta_j)\Phi_i(\theta_j)}{\sum_{j \subset k} \phi(\theta_j)}, \qquad [4]$$

where θ_j is the median of the j^{th} quantile (j = 1, 2, ..., 120) of the grade-point-average distribution (ϕ) and Φ_i is the ICC ogival function for item i. Comparison of a number, P, chosen randomly from a uniform distribution over the interval between zero and one with the difficulty, $\Phi_i(\theta_j)$, of item i for a student having grade-point average θ_j determined the response of the student to the item: correct if

Figure 1
Monte Carlo Distributions of Student Grade-Point Averages
Showing 25% (Cross-Hatched) and 50% (Cross-Hatched plus Striped) Failure Rates
with Item Characteristic Curves for Which
the Mean's Percentile and the Standard Deviation Are
25 and 1 (a), 50 and 1 (b), 25 and 2 (c), and 50 and 2 (d)



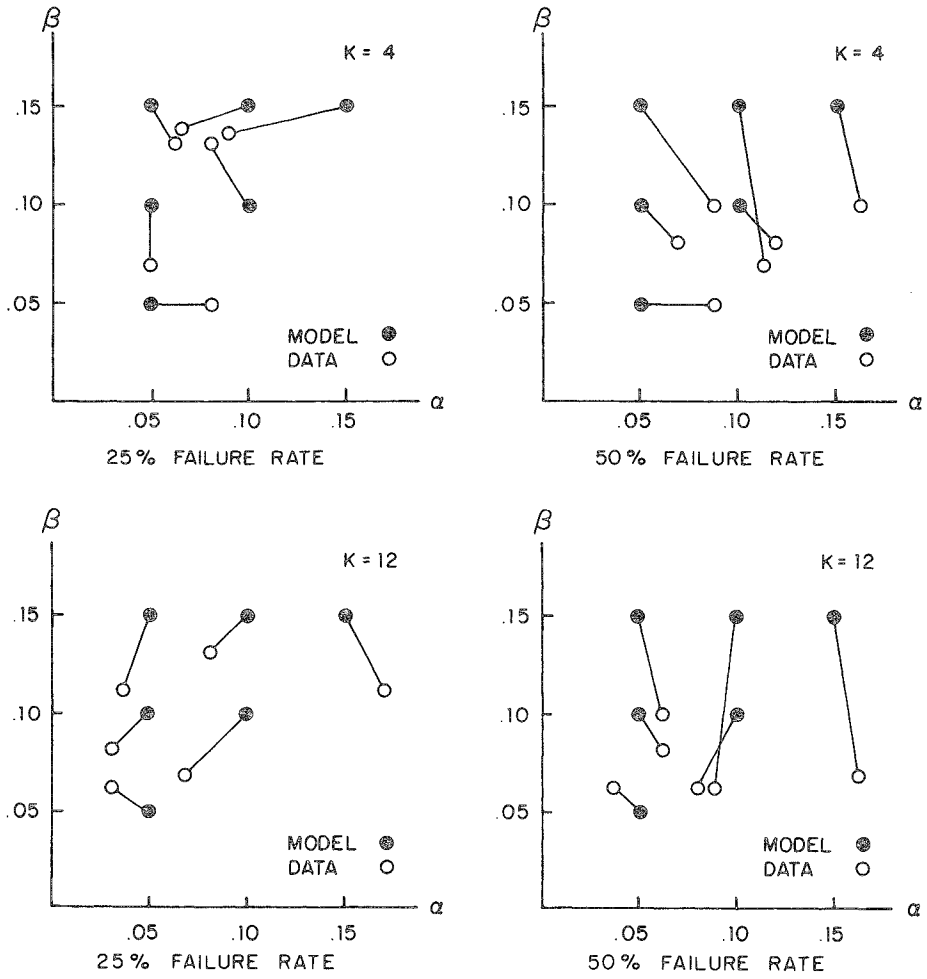
$P < \Phi_i(\theta_j)$, incorrect otherwise. The failure fraction $\int_{-\infty}^{\theta_0} \phi(t)dt$ determined the value of θ_0 separating success from failure. An acceptance error was recorded whenever L_n went above $(1 - \beta)/\alpha$ and θ_j was less than θ_0; a rejection error was recorded whenever L_n went below $\beta/(1 - \alpha)$ and θ_j was greater than θ_0. The number of acceptance errors divided by the number of students for whom θ_j was below θ_0 should tend to be equal to α; the number of rejection errors divided by the number of students for whom θ_j was above θ_0 should tend to be equal to β.

Results

Representing in succession the four ICCs shown in Figure 1, Figures 2 through 5 show the error-rate results. The two columns in each figure represent different failure fractions; the two rows, different values of K. Generally, the observed values of α and β (open circles) rather closely approximate their nominal values (solid circles). The approximation tends to be better for 2σ ICCs than for 1σ ICCs and for ICCs centered at failure fraction centiles than for ICCs centered elsewhere. No difference in goodness of approximation is readily noticeable for the two values of K.

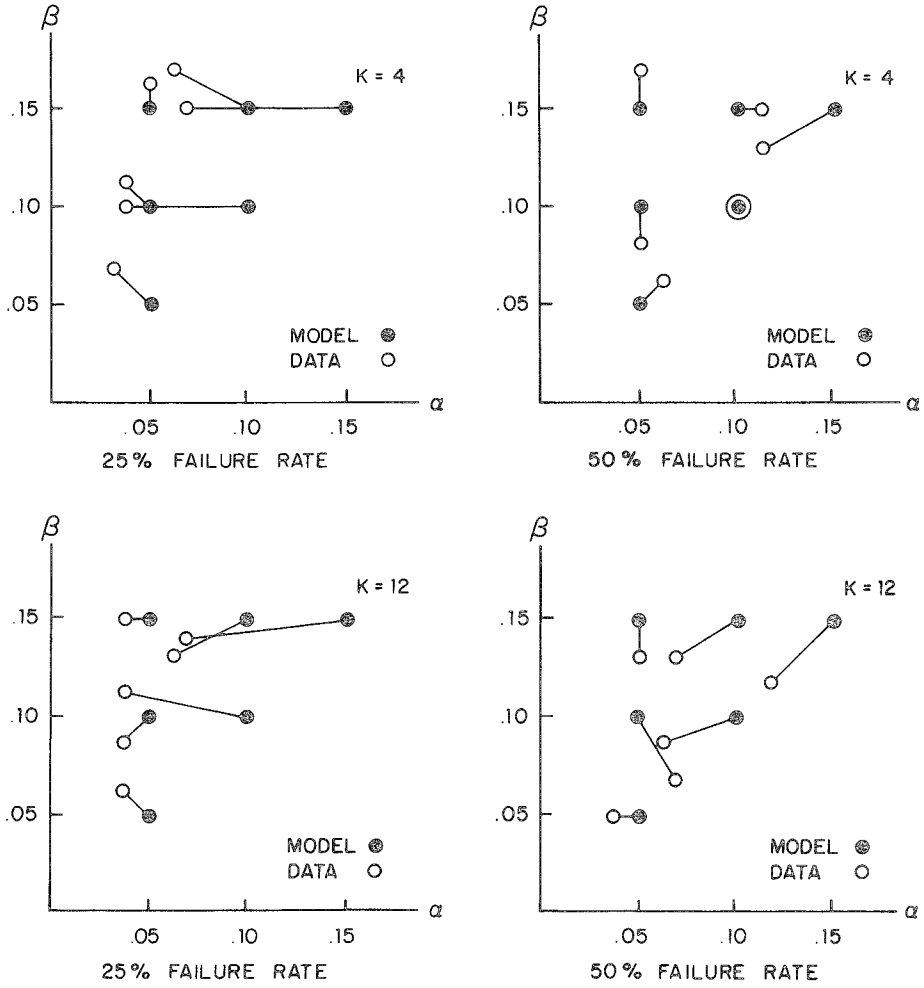
Table 1 contains the mean and standard deviation of test length for each of the 96 simulations. Both these statistics tend to increase or decrease together. Both are notably larger for $K = 12$ than for $K = 4$, for 2σ ICCs than for 1σ ICCs, and for smaller values of α and β than for larger values. All these results agree more or less with expectations except for those for the two values of K. Why should test length tend to be so much greater for $K = 12$ than for $K = 4$, especially when their approximations of α and β tend to be equally good? The answer would seem to depend on the tendency for the terms summed in L_n to be closer in value to their neighbors for the larger than for the smaller value of K.

Figure 2
Comparison of Observed (Open Circles) with Preset (Solid Circles) Acceptance (α) and Rejection (β) Error Rates for 4 (Upper) and 12 (Lower) Quantiles and 25% (Left) and 50% (Right) Failure Rates in 1,000 Monte Carlo Sequential Tests Represented by Figure 1(a)



These terms, particularly, include the neighboring two on either side of the quantile boundary separating success from failure. Following each item response, the change in L_n should thus tend to be smaller, and the test length correspondingly larger, for $K = 12$ than for $K = 4$. The results generally indicate that for $K = 4$ not only may observed α and β values approximate their nominal counterparts well enough but also test length means and standard deviations may be low enough for practical use, particularly when β is no smaller than .15 and α is no smaller than .05.

Figure 3
Comparison of Observed (Open Circles) with Preset (Solid Circles) Acceptance (α) and Rejection (β) Error Rates for 4 (Upper) and 12 (Lower) Quantiles and 25% (Left) and 50% (Right) Failure Rates in 1,000 Monte Carlo Sequential Tests Represented by Figure 1(b)

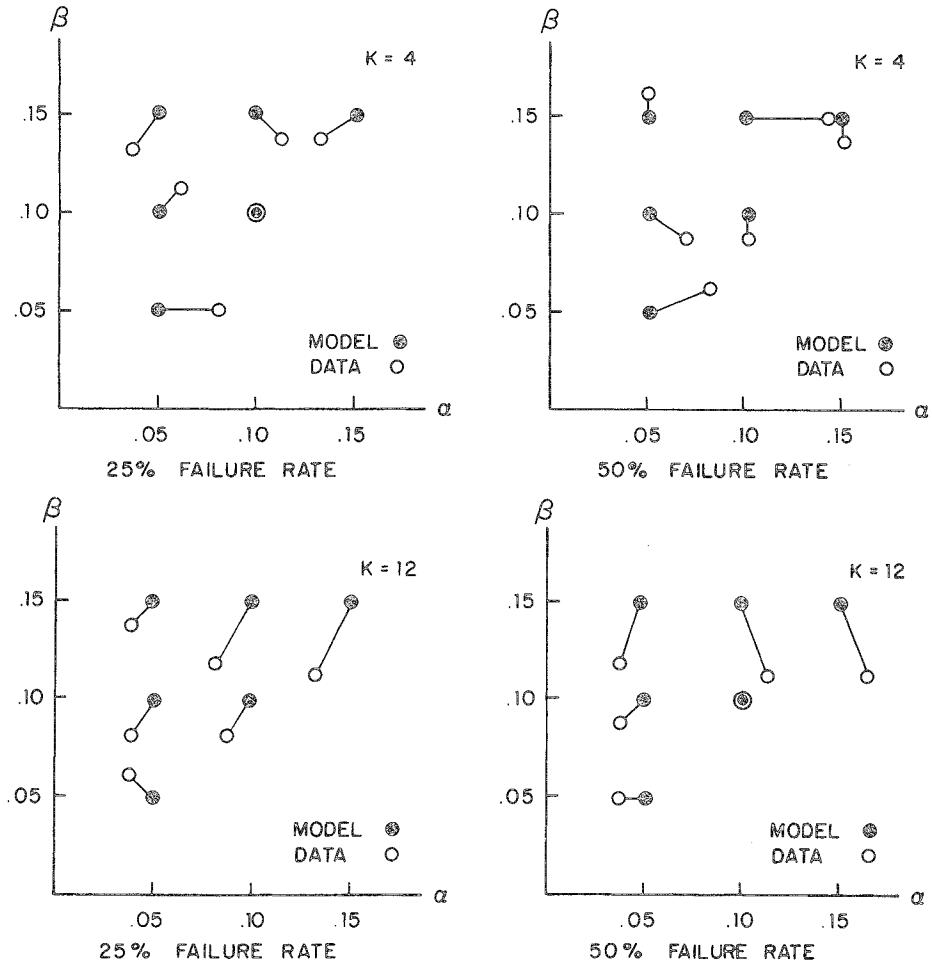


Application to Real Data

Application problems exist because groups of applicants having the same criterion scores may overlap quantile boundaries. This overlap possibility may, in fact, be responsible for two application problems. One is the determination of p_{ik} values, and the other is the interpretation of error rates. Monte carlo simulation avoided these problems by using a continuous criterion distribution.

The problem involved in the determination of a p_{ik} value when overlap exists is how to assign members of a group that overlaps the boundary between quantile k and one of its neighbors ($k - 1$ or $k + 1$). The most straightforward solution to this problem would seem to be to compute p values

Figure 4
Comparison of Observed (Open Circles) with Preset (Solid Circles)
Acceptance (α) and Rejection (β) Error Rates for 4 (Upper) and 12 (Lower) Quantiles
and 25% (Left) and 50% (Right) Failure Rates
in 1,000 Monte Carlo Sequential Tests Represented by Figure 1(c)



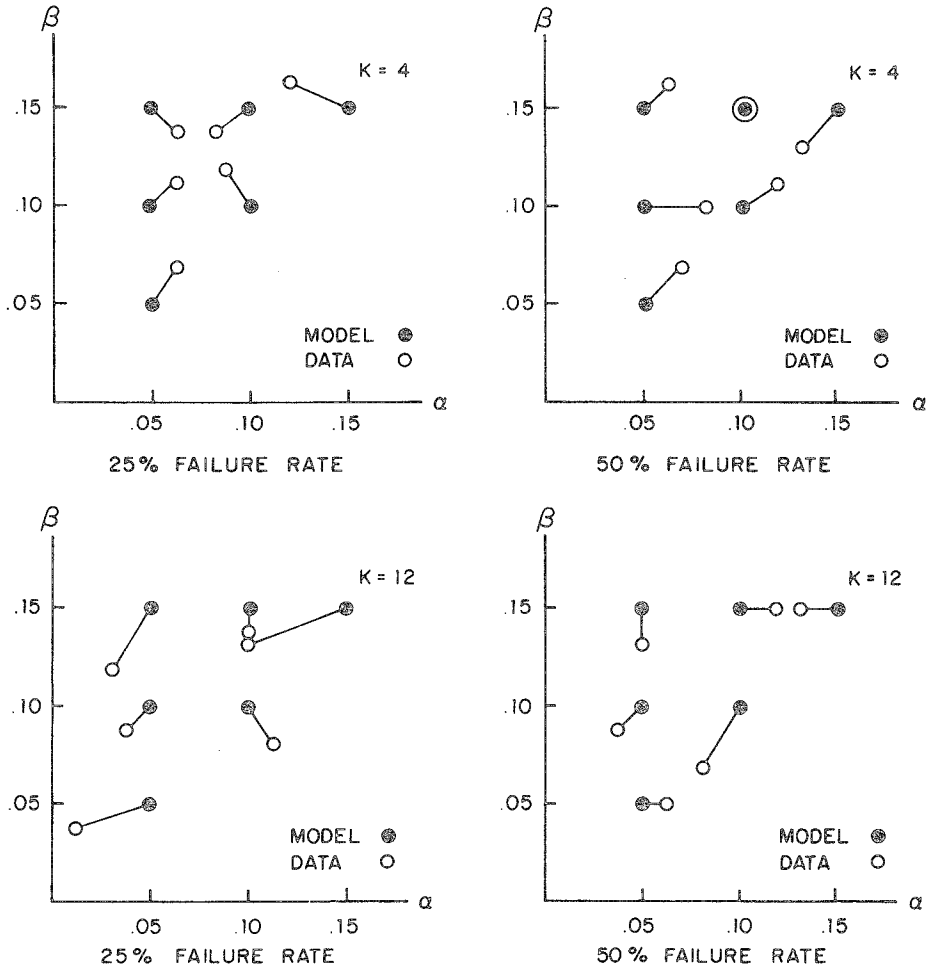
separately for the nonoverlapping and overlapping groups spanning quantile k and then to weight the contributions to p_{ik} of these groups by their proportional representation in the quantile:

$$p_{ik}' = \frac{N_L p_{iL} + N_M p_{iM} + N_H p_{iH}}{N_L + N_M + N_H}, \qquad [5]$$

large enough group may overlap both boundaries of the quantile to make N_M equal to zero and N_L and N_H the same N.

The interpretation of error rates may be a problem because of possible overlap of the boundary separating success from failure on the criterion. Does a member of the overlapping group count as an

Figure 5
Comparison of Observed (Open Circles) with Preset (Solid Circles) Acceptance (α) and Rejection (β) Error Rates for 4 (Upper) and 12 (Lower) Quantiles and 25% (Left) and 50% (Right) Failure Rates in 1,000 Monte Carlo Sequential Tests Represented by Figure 1(d)



where the N's are the numbers of individuals in the quantile and the p's the proportions of correctly responding individuals from the low (L) overlapping group, the middle (M) nonoverlapping group or groups, and the high (H) overlapping group. Any (but not all) of the N's may be equal to zero, and a error if accepted or rejected? What kind of error, acceptance or rejection? Proportional representation also provides a solution to this problem. If the overlapping group consists of five individuals with two supposed to be on the lower and three on the upper side of the boundary separating success from failure, for example, then the number of the acceptance and rejection errors to be recorded for the group depends on the number of applicants actually accepted or rejected, as shown in Table 2. The failure fraction that requires a certain number of individuals in the overlapping group to be accountable as failures thus effectively classifies rejections in excess of this number as rejection errors and rejections short of this number as acceptance errors.

Table 1
Mean and Standard Deviation of Test Length in 96 Monte Carlo Studies

Number of Quantiles and Error Probability		Item Characteristic Curve Mean and Standard Deviation													
		25% Failure Rate								50% Failure Rate					
Accept	Reject	25%ile				50%ile				25%ile				50%ile			
Bad	Good	1σ		2σ		1σ		2σ		1σ		2σ		1σ		2σ	
4 Quantiles																	
.05	.05	10	12	37	52	12	15	38	50	16	22	47	67	12	17	45	57
.05	.10	9	9	27	33	9	11	27	35	12	17	35	48	10	12	36	58
.05	.15	7	8	24	27	7	7	21	26	11	14	30	41	8	10	27	38
.10	.10	6	6	22	26	7	9	22	27	8	10	27	40	8	9	27	37
.10	.15	6	6	17	19	6	7	17	22	8	10	22	31	6	9	22	33
.15	.15	5	5	14	15	5	5	14	17	6	6	17	23	5	6	16	23
12 Quantiles																	
.05	.05	19	41	76	196	21	58	62	134	33	91	107	279	32	83	107	290
.05	.10	13	32	40	98	14	35	43	97	21	51	68	216	20	53	60	180
.05	.15	9	18	35	89	9	21	33	83	18	41	46	136	13	29	43	102
.10	.10	11	27	33	87	11	26	34	80	15	31	50	142	14	38	39	93
.10	.15	7	14	22	43	7	18	21	48	11	26	33	96	9	24	32	91
.15	.15	4	7	19	40	6	10	16	31	9	22	23	63	6	12	23	50

Note. Item characteristic curve mean is in form equivalent grade-point-average percentile of student who has .50 probability of answering item correctly; unit (σ) is standard deviation of student grade-point-average distribution.

Use of these treatments of overlap on real data demonstrated that selective testing, which worked on monte carlo data, can have corresponding success in practical applications.

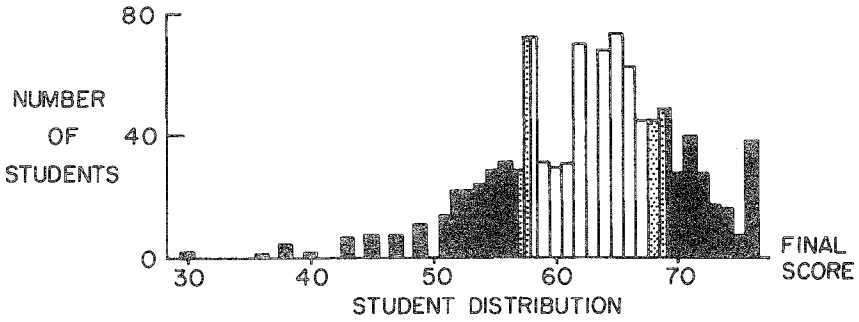
Method

The real-data pool consisted of the responses (correct/incorrect) of 960 Navy enlisted men to the 70 items of the Navy Electronics Technician Selection Test (ETST) together with the final numerical grades of these men in the Navy Basic Electronics and Electricity School in San Diego, California. The histogram in Figure 6 describes the frequency distribution of these grades. The shaded, stippled, and blank areas represent the different failure rates used in eight separate studies: 25% and 75% with $K = 4$ and 20% and 80% with $K = 5$. Use of the overlap treatments just described resolved the problems arising from the overlap apparent in a number of the score groups. Two studies used each failure rate; students unsorted by Item 70 were accepted in one and rejected in the other. Although each en-

Table 2
Errors Recorded for Members of a Criterion Score Group Containing Two Failures and Three Successes as a Function of the Number Rejected

Number Rejected	Acceptance Errors	Rejection Errors
0	2	0
1	1	0
2	0	0
3	0	1
4	0	2
5	0	3

Figure 6
Real-Data Distribution of Final-Examination Scores for 960 Students Showing 20% (Left Solid), 25% (Left Solid plus Stippled), 75% (Complement of Right Solid plus Stippled), and 80% (Complement of Right Solid) Failure Rates



listed man took the entire 70-item test, computer runs simulated the sequential procedure by selecting one item at a time. The order of selection corresponded to the ranking of the correlations between item responses and final grades. The correlation between the entire test and the final-grade criterion was .60.

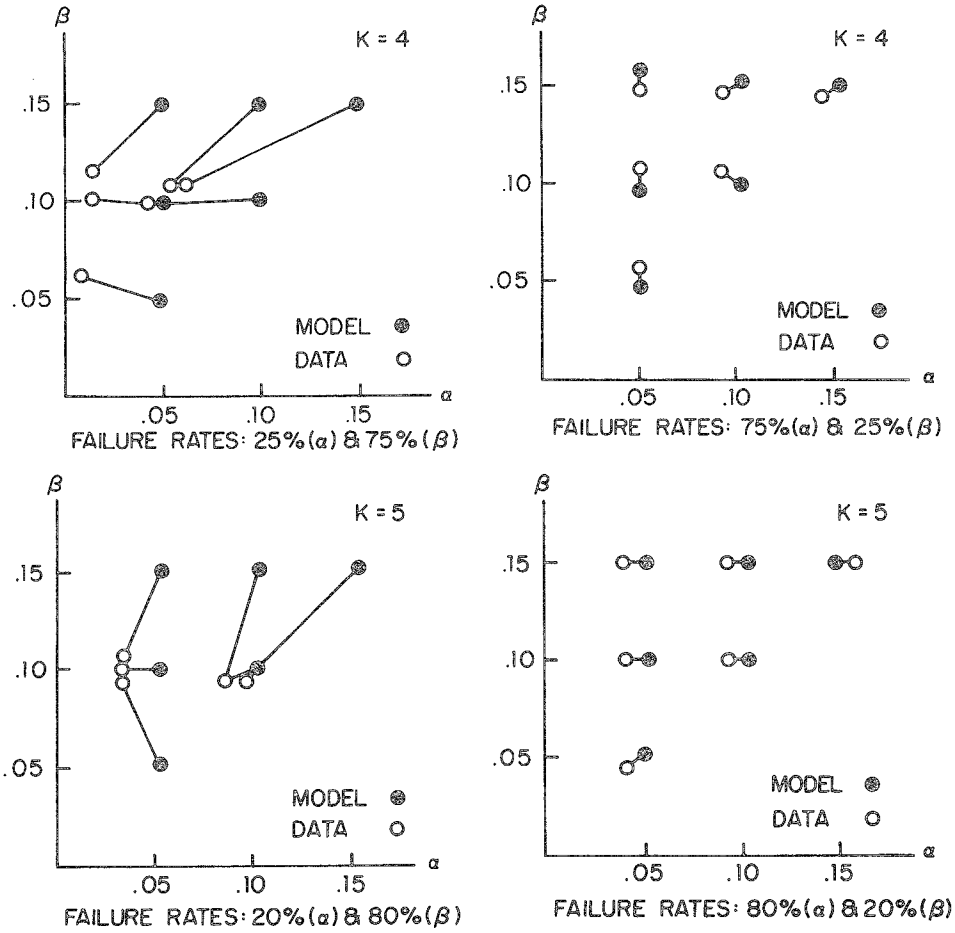
Results

Figure 7 shows the error-rate results. The upper two graphs in the figure compare the expected (solid circles) and observed (open circles) error rates for $K = 4$; the lower two graphs make the same comparisons for $K = 5$. In each comparison, the failure rates differed for the computation of the observed α and β values so that the groups used to compute the α and β coordinates of the open circles consisted of 240 (upper) or 192 (lower) men in the left graphs and 720 (upper) or 768 (lower) in the right graphs. For students unsorted by Item 70, the acceptance and rejection decisions worked in the left graphs to inflate and in the right graphs to deflate the observed α and β values. Aside from this slight inflationary or deflationary effect, the accuracy and stability of these values depend on the sizes of the groups used to compute them. The difference in accuracy and stability between the left and the right graphs reflects this dependence. The closeness of the observed to the expected values in the right graphs indicates that the corresponding discrepancies in the left graphs are due largely, if not entirely, to sampling error. Selective testing thus appears to work on real as well as on monte carlo data.

Table 3 presents the mean test lengths (left cell entries) and 70-item frequencies (right cell entries) obtained in the real-data studies. Resembling the monte carlo results presented in Table 1 for the 1σ items, the means here range from 5 for $\alpha = \beta = .15$ with a large failure rate to 17 or 19 for $\alpha = \beta = .05$ with a small failure rate. Each combination of α and β has only 4, rather than 8, test length means and 70-item frequencies because these values do not depend, as the observed α and β values do, on the acceptance or rejection of a student unsorted by Item 70. The mean test lengths tended to be well below 70, and in most cases relatively few of the 960 students required as many as 70 items for a selection decision.

These results support the supposition that sequential tests are more efficient than their conventional counterparts. A direct conventional-sequential comparison strengthened this support. The 70-

Figure 7
Comparison of Observed (Open Circles) with Preset (Solid Circles) Acceptance (α) and Rejection (β) Error Rates for 4 (Upper) and 5 (Lower) Quantiles with Failure Rates Equal to 25% for Observed α and 75% for Observed β (Upper Left), 75% for Observed α and 25% Observed β (Upper Right), 20% for Observed α and 80% for Observed β (Lower Left), and 80% for Observed α and 20% for Observed β (Lower Right) in Real-Data Sequential Tests of 960 Students Described by Figure 6



item ETST provided the conventional data, and one of the corresponding selective tests provided the sequential data for the comparison. Involving an 80% failure rate with $\alpha = .05$ and $\beta = .15$, the particular selective test compared had a mean length of 9 items and a 70-item frequency of 10 (see Table 3). Table 4 shows the corresponding decision-outcome percentages. The 4 in the lower right cell, for example, is the overall percentage for the accepted 5% of the 80% failures ($.04 = .05 \times .80$). The selection ratio (Ψ), represented by the marginal entry in the Accept column, is .21. The base rate (Ω), complementary to the .80 failures, is .20; without testing, this is the probability of selecting a potentially successful student. The probabilities of successful selection with testing (ω's) differ markedly, not

Table 3
Mean Test Length and 70-item Frequency for 960 Real Students

Number of Quantiles and Error Probability		Failure Rate			
		25%		75%	
α	β	Mean	Frequency	Mean	Frequency
4 Quantiles					
.05	.05	17	98	13	27
.05	.10	12	57	10	12
.05	.15	9	31	9	8
.10	.10	10	32	8	3
.10	.15	7	13	7	1
.15	.15	6	8	5	0
		Failure Rate			
		20%		80%	
		Mean	Frequency	Mean	Frequency
5 Quantiles					
.05	.05	19	127	12	20
.05	.10	15	73	11	14
.05	.15	11	44	9	10
.10	.10	11	50	8	6
.10	.15	8	18	7	4
.15	.15	7	15	5	1

only from this value, but also from each other for the sequential and conventional tests. Table 4 indicates that for the sequential test the probability of successful selection is 17/21, or .81; the Taylor-Russell tables (Taylor & Russell, 1939) indicate in the case of a .20 base rate and .20 selection ratio that for the conventional test, with its predictive validity of .60, the corresponding probability is .50. While the 70-item conventional test improved the probability of selecting a potentially successful student from .20 to .50, therefore, the improvement was substantially greater for the sequential test with its expected length of only 9 items: from .20 to .81. Although the reduction in testing time just illustrated is greater than the 50% predicted by Green (1970) and demonstrated by Linn et al. (1972), the comparison was not altogether unfair. The sequential test did not consist only of items most discriminating at θ_0 but, rather, of items that correlated maximally with the overall final-grade criterion. The conventional test might have done better if it contained only items most discriminating at θ_0, but,

Table 4
Decision-outcome Matrix for Sequential Test with 80% Failure Rate and α=.05 and β=.15

	Decision		
Outcome	Reject	Accept	Total
Success	3	17	20
Failure	76	4	80
Total	79	21	100

Note. Entries are percentages.

for the same reason, so might the sequential test. Sequential testing for selection thus compares favorably on real data with conventional testing for the same purpose.

Discussion

In a sequential, as opposed to a conventional, test the number and identity of items can vary from individual to individual. Extensively studied, adaptive tests are sequential tests whose purpose is measurement; the varying number and identity of items in an adaptive test can control the error of measurement for every individual who takes the test. Adaptive test theory involves formulaic specification of ICCs, and application of the theory requires prior estimation of the parameters of these curves. Less difficult to apply, sequential testing for the purpose of selection, as described here, involves no formulaic specification of ICCs and requires, instead of parameter estimation for each item, only estimation of the proportion of individuals who answer the item correctly in each quartile, or smaller quantile, of the criterion distribution.

Although these differences in purpose and ease of application are important, more basic is the difference in the abscissas of the ICCs used in the two forms of a sequential test. The adaptive test abscissa is a latent variable that has no direct empirical referent external to the test; in contrast, the selective test abscissa is the criterion variable on which the test is supposed to predict success or failure. The requirement of local independence on the abscissa, an assumption that defines the latent variable of an adaptive test, is thus a condition of a selective test that real data may or may not satisfy. The products in Equation 1 express this condition. The numerator and denominator of this equation contain sums representing alternative events. Each of these events is a sequence of correct or incorrect item responses by an individual whose criterion measurement lies within a specific quantile (k). The corresponding (k^{th}) product of response probabilities in Equation 1 approximates the probability of the response sequence for the individual to the extent that the condition of local independence exists over the entire quantile. The successful application of Equation 1 to real data indicates the existence of this condition in these data. The real-data results are thus empirical as well as illustrative. These results show, in corroboration of theory, not only that a selective test is capable of greater selection accuracy with fewer items than a corresponding conventional test but also that uniquely in a selective test—different from its adaptive, sequential-mastery, and conventional counterparts—this selection accuracy is under the control of the test user.

References

Armitage, P. Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society,* 1950, *12,* 137-144.

Epstein, K. I., & Knerr, C. S. Applications of sequential testing procedures to performance testing. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference.* Minneapolis MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Ferguson, R. A model for computer-assisted criterion-referenced measurement. *Education,* 1970, *91,* 25-31.

Green, B. F. Comments on tailored testing. In W. Holtzman (Ed.), *Computer assisted instruction, testing, and guidance.* New York: Harper & Row, 1970.

Kalisch, S. J. A model for computerized adaptive testing related to instructional situations. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Kingsbury, G. G., & Weiss, D. J. A comparison of ICC-based adaptive mastery testing and the Wald-

ian probability ratio method. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

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

Lord, F. M. *Some test theory for tailored testing.* (Research Bulletin RB-68-38). Princeton NJ: Educational Testing Service, 1968.

Reckase, M. Some decision procedures for use with tailored testing. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Taylor, H. C., & Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology,* 1939, *23,* 565-578.

Wald, A. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics,* 1945, *16,* 117-186.

Weiss, D. J. (Ed.). *Proceedings of the 1977 Computerized Adaptive Testing Conference.* Minneapolis MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Weiss, D. J. (Ed.). *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Acknowledgments

This research was supported by the Naval Postgraduate School Research Foundation with funds provided by the Chief of Naval Research.

Author's Address

Send requests for reprints or further information to R. A. Weitzman, Department of Administrative Sciences, Naval Postgraduate School, Monterey CA 93940.

ERROR CORRECTION

Weitzman, R. A. *Sequential testing*, Volume 6, Number 3, pp. 337–351.

Due to a typesetting error, several lines at the bottom of page 344 were interchanged with lines on page 345.

The lines following Equation 5 on page 344 should read as follows:

where the N's are the numbers of individuals in the quantile and the p's the proportions of correctly responding individuals from the low (L) overlapping group, the middle (M) nonoverlapping group or groups, and the high (H) overlapping group. Any (but not all) of the N's may be equal to zero, and a large enough group may overlap both boundaries of the quantile to make N_M equal to zero and N_L and N_H the same N.

The text below Figure 5 on page 345 should read as follows:

The interpretation of error rates may be a problem because of possible overlap of the boundary separating success from failure on the criterion. Does a member of the overlapping group count as an error if accepted or rejected? What kind of error, acceptance or rejection? Proportional representation also provides a solution to this problem. If the overlapping group consists of five individuals with two supposed to be on the lower and three on the upper side of the boundary separating success from failure, for example, then the number of the acceptance and rejection errors to be recorded for the group depends on the number of applicants actually accepted or rejected, as shown in Table 2. The failure fraction that requires a certain number of individuals in the overlapping group to be accountable as failures thus effectively classifies rejections in excess of this number as rejection errors and rejections short of this number as acceptance errors.

Readers should remove this page and insert it in their copy for future reference.