Use of Sequential Testing to Prescreen Prospective Entrants into Military Service

R. A. Weitzman
Naval Postgraduate School

The objective of this research was to study the possible recruiting station use of a form of sequential testing called selective testing to prescreen applicants for military enlistment.

In selective testing (described by Weitzman, 1982), an applicant responds at a computer terminal to one item at a time until the totality of his/her responses indicates either an acceptance or a rejection decision with preset error probabilities: α, the probability of accepting an applicant who will fail, and β, the probability of rejecting an applicant who would succeed if accepted. Although different applicants generally respond to different numbers of items, the average of these numbers tends to be small (less than 20), primarily depending on the magnitudes of the preset error probabilities. The validity of the selection decision requires that successive items be uncorrelated for applicants who have equal values of the performance variable (the criterion) that the test is used to predict. This local independence requirement was evidently met in a previous application of the method involving 960 Navy enlisted men who had taken both an entrance and a final examination for a technical training course. Application of the method to the entrance examination to predict passing or failing on the final examination resulted in observed error proportions that closely matched preset error probabilities (Weitzman, 1982).

The first use of sequential testing to classify individuals was an application to dichtomous classification by Linn, Rock, and Cleary (1972) of a sequential procedure developed by Armitage (1950) for polychotomous classification. This procedure used the value of an objective function to determine when testing should be terminated. Though related monotonically to this value, the observed rates of classification errors were not subject to control by the procedure.

Selective testing, a form of sequential testing, can both concentrate its accuracy at the cutting score and control the probabilities of selection errors. Selective testing is an adaptation 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). The mastery decision in each of these adaptations tends to have error rates that are no higher than preset values only for students whose subject matter mastery corresponds to proportions that fall outside an indifference region that the test user must specify.

Selective testing, by contrast, works to control the error rates for everyone. This control requires monitoring a probability-ratio test statistic, computed after each item response, to determine whether the statistic has reached a value farther from one than an upper or lower critical value. Testing continues until the test statistic has reached one or the other of these two critical values.

Before testing, a standardization group of applicants is divided into K quantile groups on the criterion. The test statistic is a function of the proportion (p_{ik}) of standardization group applicants within criterion quantile group $\underline{k}$ who answer item i (i = 1, 2, ..., n) correctly:

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

where K* designates the quantile group immediately above the criterion measurement separating success from failure and x_i equals 1 for a correct and 0 for an incorrect response to item $\underline{i}$. According to Wald (1945), the critical values for L_n are $(1 - \beta)/\alpha$ for an acceptance and $\beta/(1 - \alpha)$ for a rejection decision.

Method

The data consisted of the responses (correct/incorrect) of 1,020 Navy recruits to 200 items of the Armed Services Vocational Aptitude Battery (ASVAB) together with the scores of these recruits on the Armed Forces Qualification Test (AFQT), which functioned as the criterion. The AFQT is actually a composite of four of the ASVAB components: Arithmetic Reasoning (AR), Paragraph Comprehension (PC), Word Knowledge (WK), and Numerical Operations (NO). The 200 ASVAB items used as predictors represented all these components except NO. Table 1 shows the correlations among the AFQT and the eight components of the ASVAB represented by the items used here. In addition to AR, PC, and WK, these components were General Science (GS), Automotive-Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI).

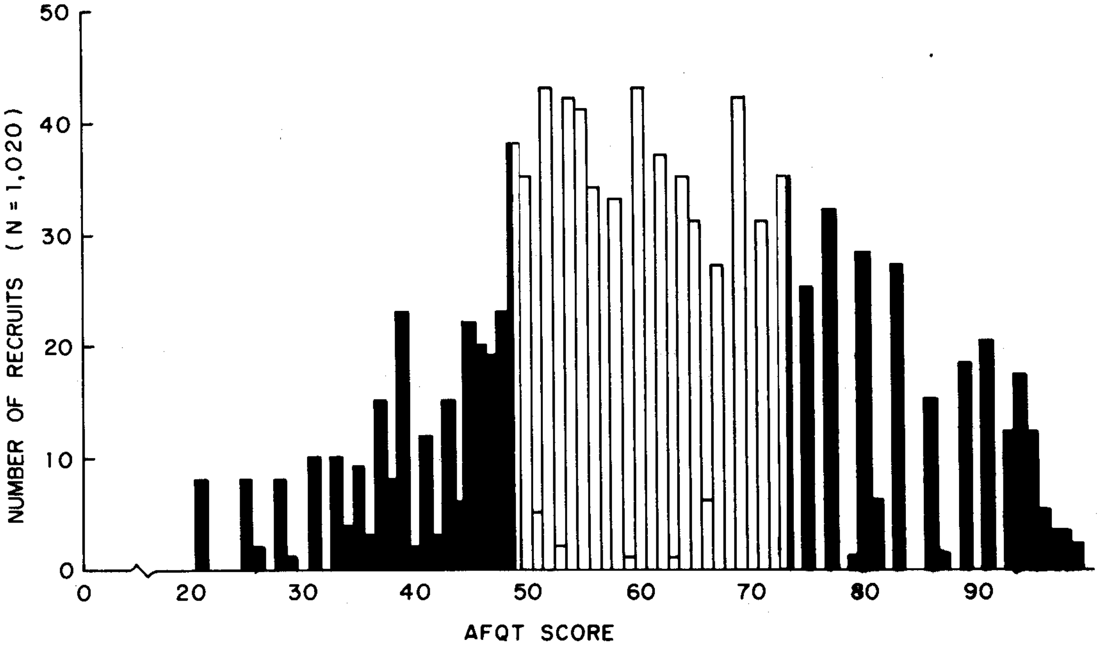
The goal of this research was to predict from a recruit's responses to the ASVAB items whether the recruit would pass the AFQT. The entire group of 1,020 recruits functioned as both the standardization and the applicant group. The histogram in Figure 1 describes the frequency distribution of the AFQT scores for these recruits. The shaded and blank areas represent the different failure rates used--25% and 75%--with the frequency distribution of AFQT scores divided into quartiles for the determination of the p_{ik} values required to compute the test statistic. Use of the overlap treatments described by Weitzman (1982) resolved the problems arising from the overlap apparent in the two boundary score groups.

Table 1
Correlations among AFQT and ASVAB Tests (N=1,032)

					Test				
Test	AFQT	GS	AR	WK	PC	AS	MK	MC	EI
AFQT		.67	.69	.66	.52	.51	.63	.59	.58
GS			.54	.71	.54	.59	.54	.58	.67
AR				.54	.55	.44	.72	.55	.52
WK					.64	.49	.48	.49	.64
PC						.44	.49	.50	.53
AS							.37	.65	.65
MK								.52	.49
MC									.61
EI									

Each failure rate was used in each of three studies exemplifying three methods of item selection. Although every recruit took the entire 200-item test battery, computer runs simulated the sequential procedure by selecting one item at a time. In two of the three studies the order of item selection corresponded directly to the ranking of the correlations between item responses and AFQT scores. In the first study the correlation was a point-biserial coefficient (Method 1); in the second, it was a phi coefficient, with AFQT scores dichotomized at the failure-rate centiles to maximize item discriminability (Method 2).

Figure 1
Frequency Distribution of AFQT Scores
for 1,020 Navy Enlisted Men Showing Failure Rates of
25% (Left Solid) and 75% (Complement of Right Solid)



Selective testing assumes local independence on the AFQT. To select items that most nearly met this assumption, the third study used as an objective function for each candidate item the ratio of the largest partial correlation between the candidate item and each item already selected, controlling for the AFQT, to the point-biserial coefficient used in Method 1. The candidate item selected was the one for which this ratio was smallest (Method 3).

The original intention in all three studies was to truncate the test at item 75. A problem arose that tended to reduce this number, however. This problem was the occurrence of p_{ik} values equal to zero or one that prevented the calculation of the test statistic for two of the 75 items selected for use in each study. Elimination of these two items thus resulted in the truncation item number actually used: 73 in all three studies.

Results

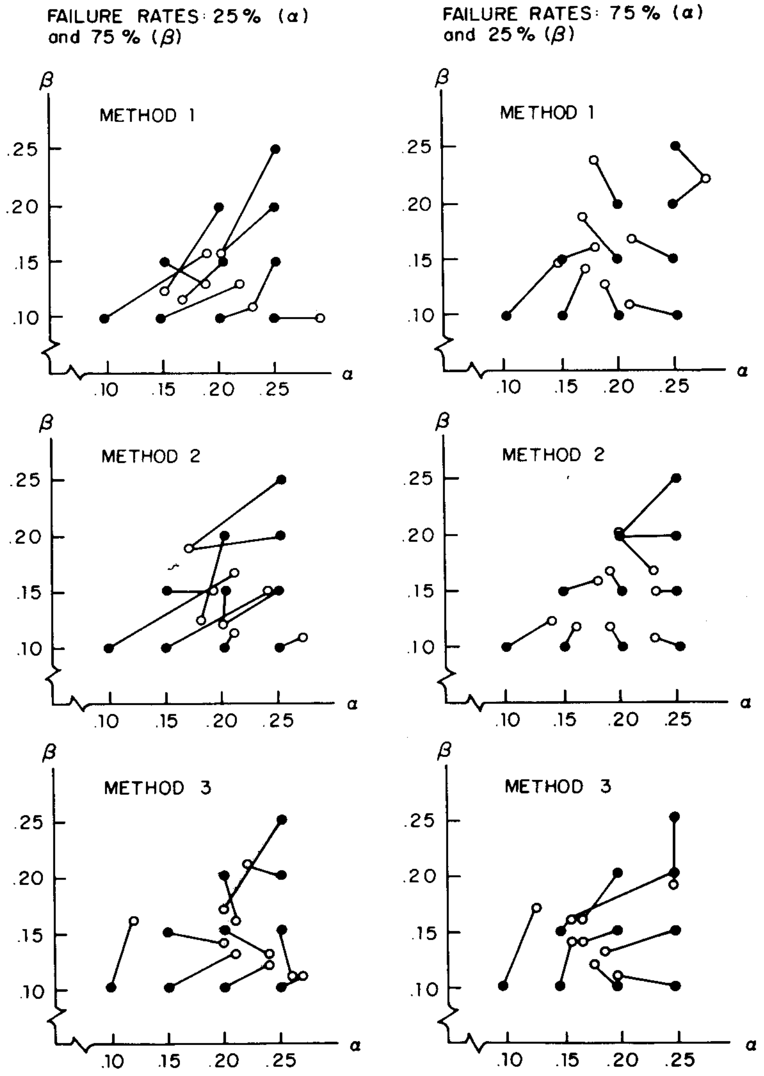
Figure 2 shows the error-rate results. The two graphs in each row compare the expected (solid circles) and observed (open circles) error rates. The three rows represent the three methods of item selection. In each comparison the failure rates differed for the computation of the observed α and β values so that the groups used to compute the α and β coordinates of the open circles consisted of 255 recruits in the left graphs and 765 recruits in the right graphs. The accuracy and stability of the observed α and β values depend on the sizes of the groups used to compute them. The difference in accuracy and stability between the left and right graphs reflects this dependence. The closeness of the observed to the expected values in the right graphs are due largely, if not entirely, to sampling error. Of the three methods, the accuracy appears best for Method 2.

Table 2 presents the mean test lengths (left cell entries) and 73-item frequencies (right cell entries) obtained in all three studies. For Method 2 the means ranged from 3 for $\alpha = \beta = .25$ with the 75% failure rate to 10 for $\alpha = \beta = .10$ with the 25% failure rate. The mean test lengths tended to be well below 73 (the maximum test length), and in no case did more than 19 of the 1,020 recruits require as many as 73 items for a selection decision. For Method 1 and Method 3 both the means and the 73-item frequencies tended to be larger than for Method 2.

Sequential tests are supposed to be more efficient than their conventional counterparts. The results just reported support this supposition. A direct conventional-sequential comparison strengthened this support. The 30-item Arithmetic Reasoning (AR) component of the ASVAB provided the conventional data, and one of the corresponding Method 2 selective tests provided the sequential data for the comparison. Involving a 25% failure rate with $\alpha = \beta = .10$, the particular selective test compared had a mean length of 10 items and a 73-item frequency of 19 (see Table 2). Table 3 shows the corresponding decision-outcome percentages. The 2.5 in the lower-right cell, for example, is the overall percentage for the accepted 10% of the 25% failures ($.025 = .10 \times .25$). The selection ratio, represented by the marginal entry in the Accept column, is .70.

The base rate, complementary to the .25 failures, is .75; without testing,

Figure 2
Comparison of Observed (Open Circles) and Preset (Solid Circles) Acceptance (α) and Rejection (β) Error Rates for Three Methods of Item Selection with 25% and 75% Failure Rates, as Shown



this is the probability of selecting a potentially successful recruit. The probabilities of successful selection with testing differ markedly, not only from this value, but also from each other for the selective and conventional tests. Table 3 indicates that for the selective test the probability of successful selection is 67.5/70, or .96; the Taylor-Russell tables (Taylor & Russell, 1939) indicate by interpolation in the case of a .75 base rate and .70 selection ratio that for the AR test, with its predictive validity of .69, the corresponding probability is .88. Although the 30-item conventional test improved the probability of selecting a potentially successful recruit from .75 to .88, therefore, the improvement was notably greater for the selective test with

Table 2
Mean Test Length (Rounded) and Truncation-Item Frequency
for 1,020 Navy Recruits

Error Probability		Failure Rate			
		25%		75%	
α	β	Mean	Frequency	Mean	Frequency
Method 1					
.10	.10	12	34	10	15
.15	.10	11	26	8	7
.20	.10	10	18	6	1
.25	.10	7	9	5	1
.15	.15	8	8	5	2
.20	.15	6	5	5	0
.25	.15	4	2	4	0
.20	.20	4	1	5	0
.25	.20	3	0	2	0
.25	.25	3	0	2	0
Method 2					
.10	.10	10	19	8	8
.15	.10	8	9	6	6
.20	.10	7	8	5	1
.25	.10	7	5	4	1
.15	.15	6	4	5	1
.20	.15	5	3	4	0
.25	.15	5	2	4	0
.20	.20	5	3	4	0
.25	.20	4	1	3	0
.25	.25	4	0	3	0
Method 3					
.10	.10	20	50	15	25
.15	.10	16	30	13	11
.20	.10	13	16	11	8
.25	.10	10	10	10	4
.15	.15	14	17	11	6
.20	.15	10	7	10	4
.25	.15	8	4	9	2
.20	.20	9	1	8	1
.25	.20	7	0	5	0
.25	.25	5	0	4	0

its expected length of only 10 items: from .75 to .96. The contrast among the different selection procedures is even sharper in terms of failure, as opposed to success, probabilities. In these terms the 30-item conventional test reduced the probability of selecting a recruit who would fail from .25 to .12, while the reduction for the selective test, with its expected length of only 10 items, was from .25 to .04. Sequential testing for selection thus compares favorably on real data with conventional testing for the same purpose.

Table 3
Decision-Outcome Percentages for Selective Test with 25% Failure Rate and $\alpha = \beta = .10$

Outcome	Decision Reject	 Accept	Total
Success	7.5	67.5	75
Failure	22.5	2.5	25
Total	30	70	100

Discussion

Methods 1 and 2 appear to produce good matches of observed with expected error rates for values of α and β between .10 and .20 (see Figure 2). Discrepancies tend to appear for values larger than .20. The observed values produced for $\alpha = .25$ or $\beta = .25$ tend to approximate the observed values for $\alpha = .20$ or $\beta = .20$. One reason for this tendency may be that the mean test lengths both for $\alpha = .20$ and $\beta = .20$ and for $\alpha = .25$ and $\beta = .25$ are approximately equal, both being nearly as small as possible, so that for both pairs of expected error rates testing tends to end at about the same item number with about equal observed error rates. Another reason is that for $\alpha = \beta = .20$ and for $\alpha = \beta = .25$, for example, the corresponding critical values tend not to differ very much: .25 and 4 for $\alpha = \beta = .20$ and .33 and 3 for $\alpha = \beta = .25$. Discrepancies also tend to occur for $\alpha = .05$ or $\beta = .05$.

Though not shown in Figure 2, because the long intersecting lines would confuse the figure, the discrepancies can be quite large. For $\alpha = \beta = .05$, in the case of a 25% failure rate, for example, the corresponding Method 2 observed values were .18 and .14. Discrepancies as large as these may be due to the large truncation item frequencies typical of low expected error rates. In the case of the preceding example, with a mean test length of 16, the frequency for truncation item 73 was 68, much larger than the largest value (19) shown for Method 2 in Table 2.

The high mean test lengths and truncation item frequencies may indicate generally poor discriminability among the ASVAB items for predicting AFQT scores. In the study reported by Weitzman (1982), involving different predictor items and a different criterion, the matches for $\alpha = .05$ and $\beta = .05$ were considerably better than here.

Table 2 shows notably lower mean test lengths and truncation item frequencies for the 75% than for the 25% failure rate. This difference may be due to the breaks in the frequency distribution, shown in Figure 1, near the 75th centile (zero frequencies for AFQT scores of 72 and 74). In contrast, the frequencies all tend to be quite large around the 25th centile. Test length and error rate accuracy may thus depend on the criterion as well as on the test items used to predict it.

The local independence assumption accommodated by Method 3 appears to hold in Methods 1 and 2 without special accommodation. The Method 3 attempt to accommodate the local independence assumption, in fact, failed noticeably to reduce the discrepancies between the observed and the expected error rates. Sample size also appears to have had no noticeable effect on Method 3 matches between these error rates. The matches are no better on the right (N = 765) than on the left (N = 255) for Method 3 in Figure 2. The mean test lengths indicate further that Method 3 may not be as good as Methods 1 and 2. For $\alpha = \beta = .10$ with a 25% failure rate, for example, the Method 3 mean test length was 20. This mean test length compares unfavorably with the corresponding mean test lengths for Method 2 (10) and for Method 1 (12).

Altogether, the results for Methods 1, 2, and 3 indicate that no special attempt to meet the assumption of local independence is necessary. Simply a correlation coefficient appears to be adequate as an objective function; and in the choice between correlation coefficients, the phi coefficient (Method 2) seems preferable. Of the three objective functions studied, this coefficient, which maximizes item discriminability, generally yielded not only the best matches between observed and expected error rates but also the lowest mean test lengths and truncation item frequencies.

At least two recommendations would appear to follow from the results of this research. Prescreening applicants for military service optimally requires, for test security, that each applicant take a unique set of test items. The first recommendation is thus that item development aim at the creation of an item bank consisting of items that are more or less equally discriminating in the region of the anticipated AFQT cutting scores. Item selection from a bank like this can be random without affecting the accuracy or the length of the test. Because the AFQT is a linear combination of ASVAB tests, the AFQT frequency distribution has a number of breaks in it. The second recommendation is that the cutting score be at one of these breaks. Since the selection decision is most difficult for applicants closest to the cutting score, making the cut at a score that is not obtained by any examinee ought to facilitate selection.

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: 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.

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: 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 Waldian probability ratio method. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: 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.

Reckase, M. Some decision procedures for use with tailored testing. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: 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.

Weitzman, R. A. Sequential testing for selection. Applied Psychological Measurement, 1982, 6, 337-351.