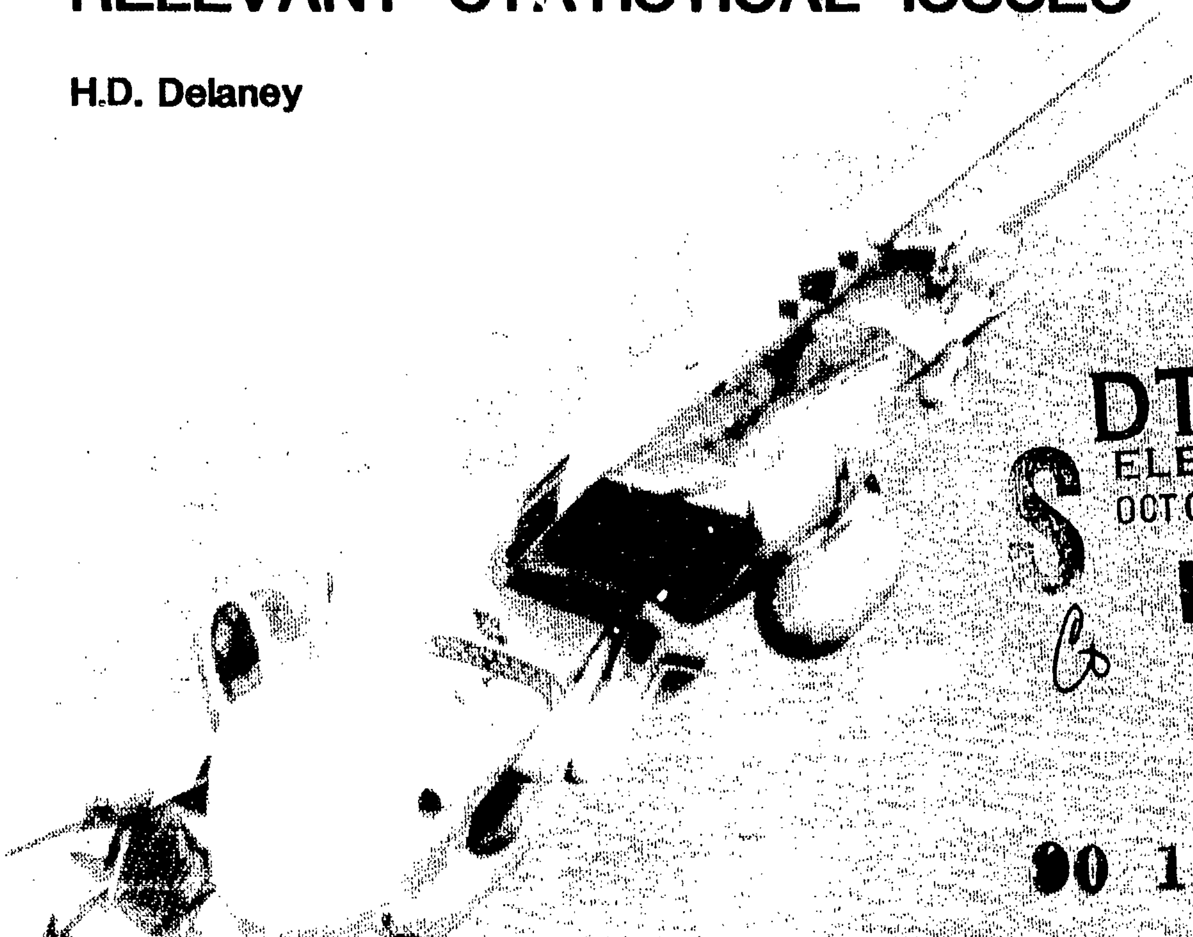
VALIDATION OF DICHOTIC LISTENING AND PSYCHOMOTOR TASK PERFORMANCE AS PREDICTORS OF PRIMARY FLIGHT TRAINING CRITERIA: HIGHLIGHTING RELEVANT STATISTICAL ISSUES

H.D. Delaney



Naval Aerospace Medical Research Laboratory
Naval Air Station
Pensacola, Florida 32508-5700

Reviewed and approved 8 July 1990

J. A. Brady

J. A. BRADY, CAPT, MSC USN
Commanding Officer

This research was sponsored by the Naval Medical Research and Development Command under Work Unit 63706N M00960.001-7007.

The views expressed in this article are those of the author and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

Volunteer subjects were recruited, evaluated, and employed in accordance with the procedures specified in Department of Defense Directive 3216.2 and Secretary of the Navy Instruction 3900.39 series. These instructions are based upon voluntary informed consent and meet or exceed the provisions of prevailing national and international guidelines.

Trade names of materials and/or products of commercial or nongovernment organizations are cited as needed for precision. These citations do not constitute official endorsement or approval of the use of such commercial materials and/or products.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1990	Interim Jun 89 - Aug 89

4. TITLE AND SUBTITLE
Validation of Dichotic Listening and Psychomotor Task Performance as Predictors of Primary Flight Training Criteria: Highlighting Relevant Statistical Issues

5. FUNDING NUMBERS
63706N
M0096 M00960.01

6. AUTHOR(S)
Harold D. Delaney*

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Aerospace Medical Research Laboratory
Naval Air Station
Pensacola, FL 32508-5700

8. PERFORMING ORGANIZATION REPORT NUMBER
NAMRL-1357

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Naval Medical Research and Development Command
National Naval Medical Center, Bldg. 1
Bethesda, MD 20814-5044

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
*U.S. Navy American Society for Engineering Education (ASEE) Summer Faculty Research Associate, summer 1989.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)
A statistical evaluation of the automated dichotic listening (DLT) and psychomotor tasks (PMT) indicated that both contributed to the prediction of primary flight training criteria. Prior to the main analyses, the extreme skewness of the DLT and PMT measures was corrected by logarithmic transformations, and between-squadron differences in flight grades were removed by transformations based on z-scores. Primary flight grades were highly correlated with the psychomotor scores (r's between -.26 and -.41) and moderately related with the dichotic listening scores (r's between -.22 and -.28). These r's were significant at an experimentwise alpha of .05. Multiple regression analysis indicated an even stronger validity coefficient when a combination of the performance measures was used (R = .442). Furthermore, the 19.5% of flight grade variance accounted for by the performance-based tests was largely independent of the 16.6% variance accounted for by a combination of current selection tests and demographic variables. For the pass/fail criterion, a statistically optimal combination of DLT/PMT variables, selection test scores, and demographic variables was specified that could be used to identify individuals who are relatively more likely to attrite. Classification matrices illustrate how such predictions could reduce attrition.

14. SUBJECT TERMS
aviation selection, psychomotor tests, dichotic listening, validation, flight training, attrition

15. NUMBER OF PAGES
24

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std Z39-18
298-102

SUMMARY PAGE

THE PROBLEM

Deriving optimal procedures for selecting pilots is a long-standing problem. The high cost of training pilots and the facts of attrition make this an important problem to address. Statistically derived combinations of predictors, including scores from automated dichotic listening and psychomotor tasks, have the potential to reduce aviator attrition through improved selection procedures.

FINDINGS

An evaluation of the automated dichotic listening task (DLT) and psychomotor tasks (PMT) based on 677 student naval aviators indicated that both contributed to the prediction of primary flight training criteria. Prior to the main analyses, issues raised by the distributions of both predictor and criterion variables were addressed. Error scores from the psychomotor tests were highly positively skewed, whereas the number correct on the DLT was highly negatively skewed.

Logarithmic transformations resulted in more nearly normal distributions and, more importantly, increased the strength of the linear relationships between the predictors and the criterion. Between-squadron differences in flight grade were removed by transformations based on *z*-scores.

Correlational analyses indicated that primary flight grades were highly related to the PMT test scores (*r*'s between -.26 and -.41) and moderately related to the DLT scores (*r*'s between -.22 and -.28). All these correlations were significant at an experimentwise alpha level of .05. Multiple regression analyses indicated even stronger validity coefficients for a combination of performance measures ($R = .442$). Further, the 19.5% of flight grade variance accounted for by the performance-based tests was largely independent of the 16.6% variance accounted for by a combination of current selection tests and demographic variables. Individual performance measures were not significantly related to the pass/fail criterion. In contrast, multiple regression techniques identified a combination of DLT/PMT variables, selection test scores, and demographic variables that could be used to identify individuals who are relatively more likely to attrite. Classification matrices were used to illustrate how this combination of variables could be used to bring about reductions in attrition rates.

RECOMMENDATIONS

Given the ample demonstrations of the validity of performance-based, particularly psychomotor, tests and the increased feasibility of such testing with microcomputer-based technology, the author recommends that such tests be transitioned into actual use for aviator selection in the Navy.

Acknowledgments

The author gratefully acknowledges the technical contributions of Pete Collyer in preparing the programs to administer the tasks and in integrating the criterion measures with the automated-task scores into a single data base for analysis. The author also benefited from helpful discussions with and/or comments on the manuscript provided by Dave Blower, Pete Collyer, Dan Dolgin, Ray Griffin, Tom Lambirth, Tatree Nontasak, and Douglas Taylor. The assistance of Sylvia Starling, Becky Dodson, and HM2 Solomon Eagles with task administration and manuscript preparation was also appreciated.

INTRODUCTION

The tests examined in this report have a relatively long history in pilot selection. This is particularly true of the tests of psychomotor ability, the predecessors of which were important components of the pilot selection battery used in the 1940s and 1950s (1). In fact, of the 20 printed and apparatus tests constituting the U.S. Aircrew Classification Battery in the early 1950s, the Complex Coordination Test, which required adjustments of stick and rudder controls, was found to have the highest validity coefficient (approximately .40) for predicting success in primary flight training (1,2). With the shift toward testing college students at many different locations, administrative and technical difficulties resulted in the suspension of the use of such psychomotor tests in the late 1950s (1). Advances in solid-state technology, however, prompted renewed interest in psychomotor testing. A mini-computer-based psychomotor test was developed by the Air Force in the early 1970s, and preliminary studies of a version implemented at this laboratory began approximately 10 years ago (3).

Large-scale validation studies of the Air Force's computerized psychomotor test have recently appeared (4,5). For example, Carretta (4) reports analyses of a study of 478 Air Force officer candidates administered a battery consisting of the Basic Attributes Test and the Air Force Officer Qualifying Test. As in the early studies, the finding again was that psychomotor tracking error scores were more strongly related (R approximately .25) to a pass/fail criterion than any other computerized or paper-and-pencil test in the battery (4).

Early analyses of an electromechanical version of the Complex Coordination Test at the Naval Aerospace Medical Research Laboratory (NAMRL) were similarly encouraging. An initial study that used as a criterion the composite flight grade of 147 student naval aviators (SNAs) revealed a correlation of -.31 with mean psychomotor tracking error (3). A subsequent validation study summarized in (6) used a dichotomous variable of outcome in primary flight training as the criterion: either pass ($n = 277$) or flight failure ($n = 17$). (The 31 other SNAs who attrited from primary flight training for other reasons, as well as 24 who switched to the Naval Flight Officer Program, were excluded from the analysis.) The comparison of complex coordination performance against other variables included in the performance-based battery known as DYNASTES (Dynamic Naval Aviation Selection Test and Evaluation System) again revealed that complex coordination error measures were the best predictors of flight failure (r's approximately .2). Preliminary validations of the micro-computer-based version of the Complex Coordination Test currently in use at NAMRL have been reported by Griffin (7,8) in conjunction with analyses of the Dichotic Listening Task (DLT).

Gopher and Kahneman (9) proposed using a dichotic listening test to predict success in flight training and reported validity coefficients of approximately .3 in an initial study with the Israeli Air Force. A version of the DLT was implemented at NAMRL in 1979 (10), and an initial validation with SNAs was reported in 1982 (11). The ceiling effect of the test (mean percentage correct > 98%) was noted early, and a version was tried with background noise as a means of dealing with this problem (11). Although background noise did lower the mean percentage correct (to approximately 91%), it also lowered the predictive validity of the test to nonsignificant levels (11).

Current computerized versions of the performance-based tests have included the DLT alone and in conjunction with a psychomotor task that is essentially a computerized form of the venerable Complex Coordination Test. The combination of the two tasks was thought to come close to duplicating aviator performance requirements (11) in the manner of what historically was termed a *job replica* test (2). In addition, the combination of auditory and psychomotor tasks carried the concept of assessing divided attention (which had initially motivated the DLT test) to higher levels by simultaneously requiring multiple responses to inputs in multiple sensory modalities: keypad responses to selected auditory signals and manual, that is, stick (hand) and rudder (foot pedal) or throttle (hand), responses to multiple visual inputs.

The primary sources of information on the validation of the dual DLT/PMT tasks to date are two previous NAMRL publications (7,8). Both reports provide encouraging evidence of the predictive validity of the tests, but they are limited by the relatively small sample sizes, particularly with regard to the pass/fail criterion. For example, the preliminary validation study reported by Griffin and McBride (7) was based on only 50 cases, and the correlation with the pass/fail criterion hinged on the mean scores on the performance tests of the subset of 5 individuals who attrited. Not surprisingly, when the recommended replications were carried out with somewhat larger samples (8), several of the results failed to hold up. Most notably, the dual DLT correct score that had been reported (7) to correlate in the range of .395 to .413 with a pass/fail criterion was subsequently found (8) to correlate in the range of only -.03 to .13 with the same criterion. Correlations with the more predictable criterion of primary flight grade came closer to replicating, although they were somewhat lower with a larger sample ($n = 95$) (cf. 8, Table 7A, and 7, Table 3). Results were also encouraging with a sample ($n = 95$) performing a backward version of the psychomotor task whereby movement of the CRT cursor was in the opposite direction of the stick and rudder controls (cf. 8, Table 7B). Griffin concludes his report by recommending the backward series "be administered to a large sample of student naval aviators to determine if the tests can account for additional variance in predicting flight training performance beyond that of current selection tests" (8, p. 11).

The purpose of the current report is to provide such a large-scale validation for the DLT/PMT tests. In addition, an attempt will be made to illustrate the relevance of certain statistical issues that are applicable not only to the current data, but more generally to the interpretation of validation studies of performance-based tests at NAMRL.

METHOD

SUBJECTS

The DLT/PMT tasks were performed by 677 student naval aviators after completing the academic portion of naval flight training and while awaiting the flight portion of primary training. The current report summarizes data on testing conducted during a period of over 2 years, from fall 1986 through the end of 1988. Subsequent to testing, attempts were made to obtain information on subjects who completed or attrited from primary flight training. Criterion information was available in the form of a dichotomous outcome variable (pass = 1, fail = 0) for 531 subjects, including 47 attrites. In addition, primary training flight grades were available for 495 subjects.

APPARATUS AND PROCEDURES

The DLT and the simplest form of the PMT were initially performed separately (single mode) to familiarize subjects with the tasks and then performed simultaneously (dual mode). Additional components were added subsequently to the PMT as detailed below. Subjects performed the series of tasks in the order indicated in Table 1. The tasks were controlled by an Apple IIe computer.

TABLE 1. Sequence of Psychomotor (PMT) and Dichotic Listening Tasks (DLT).

Order in sequence	Mode (single or dual)	Task description	Test time (min)[a] indiv.	Test time (min)[a] cum.
1.	Single	PMT stick	13	13
2.	Single	DLT	[b]15	28
3.	Dual	DLT, & PMT stick	6	34
4.	Single	PMT stick/rudder	17	51
5.	Dual	DLT, & PMT stick/rudder	6	57
6.	Dual	DLT, & PMT stick/rudder	6	63
[c]7.	Single	PMT stick/rudder/throttle	11	74

[a]Times indicated are approximate since they include typical times for reading instructions and brief breaks between tasks, which are subject-paced. Durations of the components of the tasks per se are indicated in the text.
[b]Final 65 subjects tested performed a shortened version of the DLT which required 4 min less testing time.
[c]Task was administered only to final 345 subjects tested.

Psychomotor Tasks

Subjects were required to maintain first one, then two, and finally three randomly displaced cursors on fixed targets on a CRT by manipulating joysticks and foot pedals. Subjects manipulated one Measurement Systems, Inc., joystick using their right hand to attempt to control the "stick" cursor, which was free to move throughout a rectangle covering approximately two thirds of the CRT screen. Specifically, the rectangle encompassed a 220 x 120 pixel portion of a screen that was approximately an 8.5-inch square with 280 x 160 addressable pixel locations. The target position of the cursor was indicated by crosshairs bisecting the rectangle, with the center point being slightly (10 pixels) to the right and slightly (10 pixels) above the center of the screen. The stick controlled the cursor in a backwards fashion, for example, moving the stick to the right caused the cursor to move to the left, and pulling the stick toward the subject caused the cursor to go up. Locally produced rudder pedals patterned after those of a Systems Research Laboratory Psychomotor test device were used to control the rudder cursor, which could move horizontally (over a 220 pixel distance) across the bottom of the screen. The pedals worked in conjunction with each other. Pushing the left pedal caused the cursor to move to the right, and pushing the right pedal caused the cursor to move to the left. Finally, another Measurement Systems joystick manipulated by the subject's left hand controlled the throttle cursor, which could move vertically (over a 120 pixel distance) on the left side of the screen. Pulling the throttle toward the subject caused this cursor to go down.

Single-task PMT sessions consisted of brief instructions on the screen, followed by a 3-min practice session and multiple 3-min test sessions with 30-s rest periods between sessions. There were two 3-min test sessions for the single PMT stick task, three for the single PMT stick/rudder task, and two for the single PMT stick/rudder/throttle task.

Psychomotor test scores were simply the cumulated total absolute errors from the target in pixels. For each time sampling of cursor position, absolute pixel errors were assessed separately along each dimension and summed across all dimensions represented in that task. Final scores were the sum over the many samplings of cursor positions for that task. While the number of time samplings was constant for all subjects performing a given task, it did vary over tasks but was not recorded. This prevents meaningful comparisons of errors across tasks (e.g., of stick errors in single vs. dual mode) but does not affect the results of primary interest, namely, how well errors subjects made on a given task correlate with the primary flight criteria.

Dichotic Listening Task

Subjects listened to two different series of letters and numbers presented simultaneously to their ears over binaural headphones at the rate of .7 s/item. Subjects were instructed which ear to attend to on each trial, first for a series of 16 pairs of letters and/or numbers (Part I), and then again for a series of 6 more pairs (Part II). A diagram of a typical trial is given in Table 2. Subjects were told to indicate the digits presented to the designated ear in the order of their occurrence. Responses were entered with the left hand on a keypad placed in front of the subjects. In each part of the trial, responses could be made while the items were being presented or during an interval of 1.4 s after the presentation of the last letter and/or number pair. A complete trial required 21 s. Five correct responses were possible on Part I and four on Part II of each trial. Test instructions were presented on the CRT and included six practice trials with standard auditory presentation of items but, in contrast to test trials, with visual feedback of the presented digits and the subject's responses as well. Finally, subjects completed three multiple-choice questions on the DLT and were asked to call the assistant for an explanation if they missed a question. After subjects completed the multiple-choice items successfully, a series of DLT trials was given. The series was 24 trials long for the first 612 subjects; 12 trials long for the final 65 subjects. Scores from subjects receiving the 24-trial version were halved so that the maximum score was 108 for all subjects. Mean scores on this 108-point scale were not significantly different across the groups getting the two versions of the task.

Dual Tasks

In the dual tasks, subjects performed a 4.5-min PMT and a 12-trial DLT simultaneously. The DLT task began 15 s after the PMT began, and it ended just before the PMT; PMT errors were recorded for the final 4 min of the task. Performance was scored in the same way as for the single tasks.

TABLE 2. Diagram of a Dichotic Listening Task Trial.

	Part I	
Heard by Ear(s):		
Left:		R S N S M Y 2 G B 7 F L 6 R L 5
Both:	"Trial" "3" "Right"[a]	
Right:		Y L <u>3</u>[b] S R <u>4</u> F Z <u>9</u> X F <u>0</u> F N <u>1</u> L
	Part II	
Heard by Ear(s):		
Left:	B F <u>4</u> <u>3</u> <u>7</u> <u>9</u>	
Both:	"Left"[a]	
Right:	G L 1 5 6 2	

[a]Target ear command.
[b]The digits that subjects should respond with are underlined.

RESULTS

DESCRIPTIVE STATISTICS ON ORIGINAL VARIABLES

Means, standard deviations, and an index of skewness (explained below) were computed for the DLT and PMT tasks (Table 3). In addition, information on selected background variables for the current sample is given in Table 4. Statistics for tests currently used by the Navy for selection, the Academic Qualifications Test (AQT) and the Flight Aptitude Rating (FAR), are presented for both raw score and stanine forms. Descriptive statistics on the criterion variables are shown in Table 5. Correlations of these variables with the DLT/PMT variables and with the selected background variables are presented in Table 6 in a form comparable to that used in previous reports.

TABLE 3. Descriptive Statistics for DLT and PMT.

Test	Mean	*SD*	Skewness
Dichotic Listening--Number Correct			
Single DLT	101.90	5.69	-4.36
Dual DLT1 (with PMT stick)	98.53	9.05	-2.51
Dual DLT2 (with PMT stick/rudder)	97.24	9.48	-2.14
Dual DLT3 (with PMT stick/rudder)	98.00	10.21	-3.23
Psychomotor Tasks--Cumulative Pixel Errors			
Single PMT stick	16995.02	16938.31	4.04
Dual PMT stick (with DLT1)	6239.69	6194.60	3.74
Single PMT stick/rudder	47143.86	34184.62	4.12
Dual PMT stick/rudder (with DLT2)	13869.72	10716.38	2.86
Dual PMT stick/rudder (with DLT3)	13295.59	11537.44	3.61
Single PMT stick/rudder/throttle[a]	37706.42	23042.32	3.15

[a]*n* = 345 for this task; other *n*'s are all approximately 675.

TABLE 4. Descriptive Statistics on Background Variables.

Variable[a]	Mean	*SD*	Skewness
Age	23.17	1.46	1.01
Previous flight hours	21.93	138.37	13.17
AQT (stanine)	5.70	1.29	.31
FAR (stanine)	7.14	1.63	-.53
AQT (raw score)	68.92	10.10	-.20
FAR (raw score)	37.91	6.60	.16

[a]*n* = 677 for age and previous flight hours, *n* = 666 for AQT/FAR scores.

TABLE 5. Descriptive Statistics on Criteria.

Criterion	Mean	*SD*	Skewness	*n*
Pass/fail	.9115	.2843	-2.8892	531
Flight grades	3.0493	.0353	.1208	495

TABLE 6. Correlations of DLT/PMT Measures and Background Variables with Criteria.

Variable	Pass/Fail[a]	Flight grades Orig.[b]	Flight grades z score[c]
DLT/PMT Measures			
Single DLT correct	-.03	.15	.17
Dual DLT1 correct	.01	.19	.20
Dual DLT2 correct	.06	.23	.25
Dual DLT3 correct	.10	.19	.21
Single PMT stick	-.10	-.25	-.25
Dual PMT stick (with DLT1)	-.01	-.27	-.27
Single PMT stick/rudder	-.09	-.30	-.29
Dual PMT stick/rudder (with DLT2)	-.02	-.31	-.32
Dual PMT stick/rudder (with DLT3)	-.04	-.29	-.29
Single PMT stick/rudder/throttle	-.10	-.19	-.14
Background Variables			
Age	-.04	-.07	-.10
Previous flight hours	.03	.12	.14
AQT (stanine)	.03	.15	.14
FAR (stanine)	.14	.23	.27
AQT (raw score)	.03	.15	.16
FAR (raw score)	.13	.26	.29

[a] n = approximately 530 for all correlations with pass/fail except for PMT stick/rudder/throttle where n = 205.
[b] n = approximately 490 for all correlations with flight grades--Original except for PMT stick/rudder/throttle where n = 193.
[c] n = approximately 480 for all correlations with flight grades--z scores except for PMT stick/rudder/throttle where n = 185.

STATISTICAL CHARACTERISTICS AND TRANSFORMATIONS OF PREDICTORS AND CRITERIA

Closer examination of both the predictor and criterion variables revealed that important statistical characteristics of these variables needed to be addressed in any appropriate analyses. These are discussed in turn, beginning with the predictors.

Predictors

One striking characteristic of the data both statistically and visually is the extreme skewness of the predictor variables. A plot of one of the PMT variables, the errors in the first dual test, illustrates the point (see Fig. 1). Although the mean is 6240, scores range up to 64,447. Similarly, only 1 case in a 1000 (0.1%) would be expected to be 3 standard deviations above the mean in a normal distribution. Here, 16 cases out of 677 (2.4%) are.

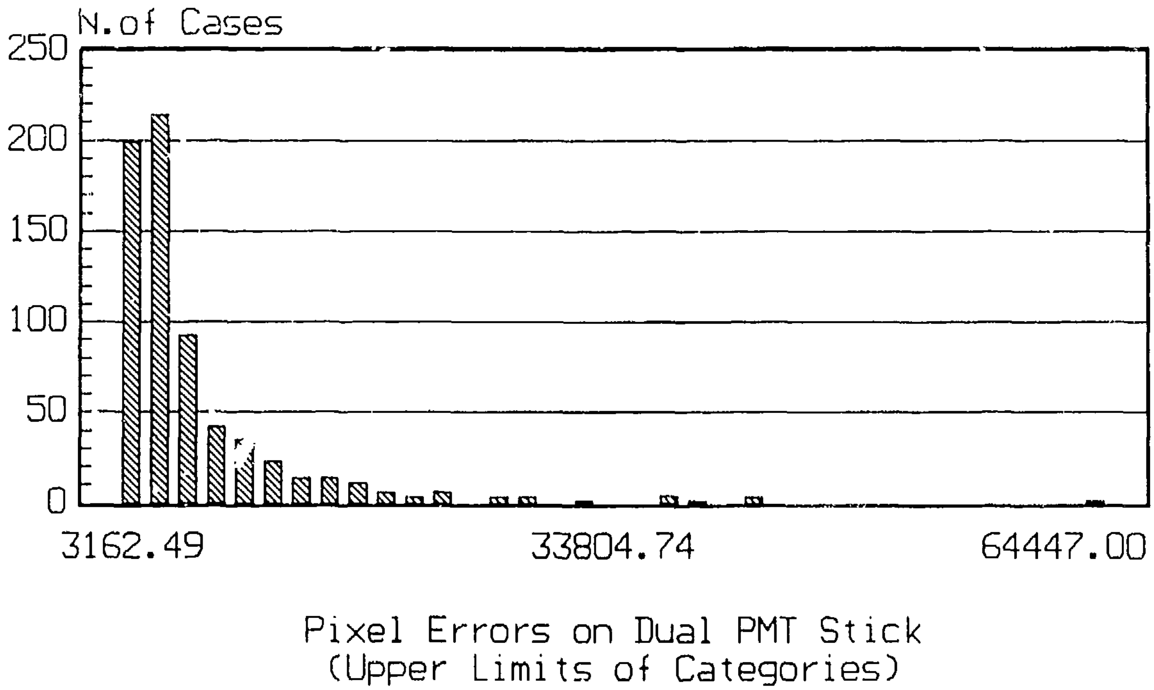


Figure 1. *Distribution of cumulative pixel errors on dual PMT stick.*

The statistical index of skewness reported is that recommended by Fisher (12), sometimes referred to as g_1. It is defined as the ratio of the third moment around the mean to the square root of the cube of the variance. The sign indicates the direction of skewness. In a normal distribution, this index of skewness is 0. Values greater than 2 are very large, and as the estimates of g_1 in Table 3 indicate, all of the DLT measures are very negatively skewed, and all of the PMT measures are very positively skewed.

Although skewness does not invalidate a variable, it can complicate the analysis of its relationship with other variables. For example, scatterplots of primary flight grade against DLT/PMT variables indicated that the skewness resulted in relationships between the predictors and the criterion that were to a certain extent nonlinear. The nonlinearity induced by the extreme scores was as follows: those scoring very poorly on the DLT/PMT tasks were worse than average in their flight grades but not as extremely low as their DLT or PMT scores would suggest. Those same outliers, because of their extremity, had the greatest influence (leverage) on the slopes of the regression lines. The presence of these extreme scores made the correlations smaller and the regression lines flatter.

Excluding these cases would have improved the situation somewhat, but less than keeping them and transforming the scores to a log scale. Log transformations resulted in a much more nearly normal distribution of the predictors (see Fig. 2) and a more linear relationship between the predictors and the flight grade criterion. Such a transformation changes the units of the scale. In the case of the PMT measures, the meaning of the units of cumulated pixel errors was not clear to begin with, so no interpretability was lost by using log cumulated pixel errors. In the case of the negatively skewed DLT scores, it was necessary first to transform the number correct to number wrong and add 1 before taking logs. This ensured that a perfect score of 108 would translate into a score of 0 on the log DLT errors scale. Also, some consolidation of variables was accomplished by first combining the scores from the two replications of the dual stick/rudder task into a single average score for the DLT as well as the PMT (e.g., dual DLT2 and dual DLT3 were averaged). Descriptive statistics for the final set of eight DLT/PMT variables are shown in Table 7.

TABLE 7. Descriptive Statistics for Transformed DLT and PMT Variables.

Transformed variable	Mean	*SD*	Skewness
Dichotic Listening--log_{10} (Number Wrong + 1)			
Single DLT	.76	.27	.11
Dual DLT1 (with PMT stick)	.88	.36	-.30
Dual DLT2,3 (with PMT stick/rudder)	.95	.30	.09
Psychomotor Tasks--log_{10} Cumulative Pixel Errors			
Single PMT stick	4.12	.29	.84
Dual PMT stick (with DLT1)	3.68	.29	.81
Single PMT stick/rudder	4.61	.21	1.01
Dual PMT stick/rudder (with DLT2,3)	4.05	.25	.76
Single PMT stick/rudder/throttle[a]	4.22	.20	.96

[a]n = 345 for this task, other n's are all approximately 675.

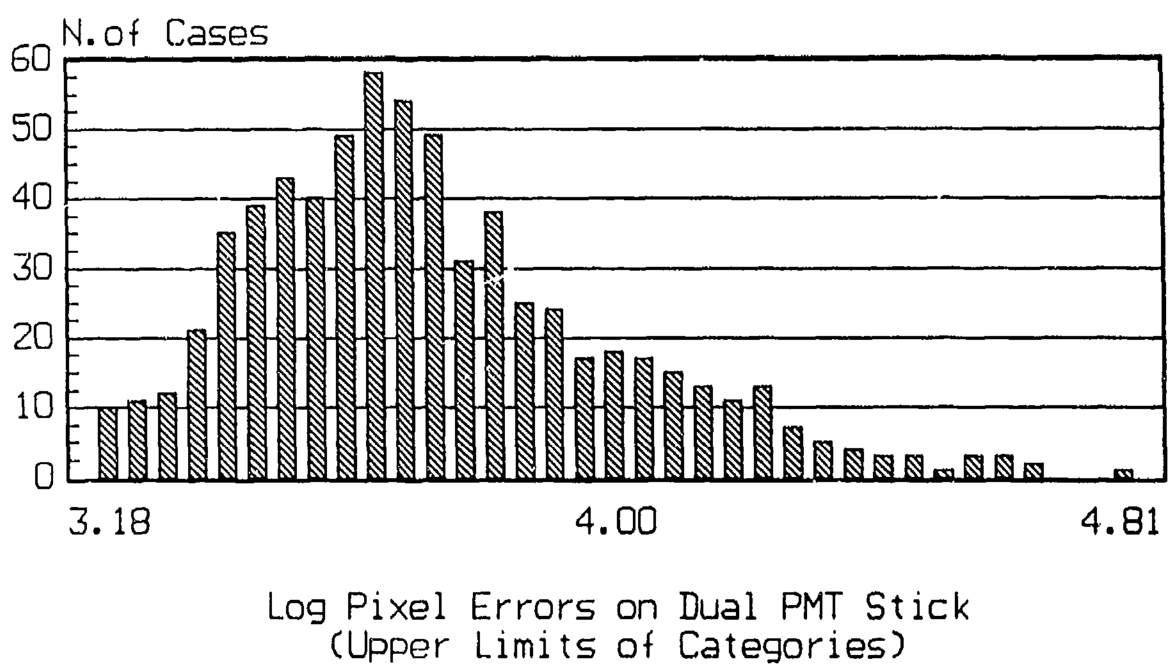


Figure 2. *Distribution of logarithm of cumulative pixel errors on dual PMT stick.*

Criteria

Prediction of primary flight training performance is complicated by statistical characteristics of the criteria as well. With regard to the pass/fail criterion, current validity coefficients do not compare with those of the World War II era primarily because of the profound effects of restriction of range (13) that results from admitting only college graduates with high AQT/FAR scores. This is even more true in the samples analyzed at the NAMRL from which attrites during School's Command have been excluded, making the proportion passing primary flight training even higher. With the pass rate at .9115, as indicated in Table 5, the maximum possible correlation of this criterion with a normally distributed variable is mathematically limited to be no greater than approximately .5 (14).

In addition, some such as Doll (15) have argued that the pass/fail criterion is inherently unpredictable because of unreliability. Doll's logic in part is that low reliability would result from either specificity (e.g., varying quotas on the number of students that can or must be graduated) or error variance (e.g., varying reasons why individuals attrite). In the current sample of 531 who completed primary flight training, the 47 attrites included only 5 flight failures, which were combined with 14 not physically qualified, 25 drops on request, and 3 academic failures. If low reliability in the pass/fail criterion does occur, it further limits the maximal correlation with any predictor. Thus, uncorrected correlations with pass/fail of .1 to .2, though accounting for a small proportion of the variance, should nonetheless be regarded as substantial given all the factors tending to depress these correlations.

The primary flight grade score, although overall approximately normally distributed and hence much more predictable than pass/fail, also had some statistical peculiarities. Students' performance in primary flight training was evaluated by one of three squadrons: VT2, VT3 or VT6. The number of students completing with each squadron and statistics on grades by squadron are shown in Table 8.

TABLE 8. Primary Flight Grades by Squadron.

Squadron	Mean	*SD*	*n*
VT2	3.04501	.0327	177
VT3	3.04504	.0321	170
VT6	3.06282	.0362	138
Overall	3.04934	.0353	485

Although at first glance such between-squadron differences in grades may appear small because of the scaling of the variable, in fact they are substantial. An analysis of variance revealed significant differences in the mean rating given by the three squadrons, $F(2, 482) = 13.879$, $p = .00002$. Follow-up Tukey tests to assess pairwise differences revealed the mean grade given by VT 6 was significantly higher ($p < .0001$) than that given by either VT 2 or 3, which in turn did not differ significantly from each other. The difference between Squadron 6's mean and that of the other squadrons amounts to half of a within-group standard deviation, certainly a nontrivial difference.

Differences due to alternative grading procedures by the three squadrons needed to be adjusted for before proceeding with other analyses. Students' grades were first converted to z scores relative to the mean and standard deviation of their squadron. Then they were rescaled so that grades for all three squadrons had means and standard deviations equal to those for the total sample, that is, 3.04934 and .0352, respectively. In this way, the specificity in a student's grade due to the squadron assigning the grade was removed from the criterion and was prevented from lessening the relationship of the criterion to the predictors. Correlations of the z-score-based transformation of the criterion with the predictors in their raw score form are shown in the rightmost column of Table 6. Note that these correlations in general are slightly larger in absolute value than those correlations with the original flight grades. Unless otherwise noted in the remainder of the report, it is this transformation that will be intended when a reference is made to flight grade.

The final correlations between the log transformed error scores and the criteria are presented in Table 9. Correlations with the two primary criterion variables of pass/fail and the z-score-based form of flight grades were tested for significance. Given there were 8 DLT/PMT scores and 6 background variables, the resulting 28 tests were required to be significant at $.05/28 = .0018$ in order to maintain experimentwise alpha at .05 (16). This in turn implied a critical r value of approximately .14 for tests based on 500 or more cases, approximately .23 for tests based on 180 cases. Even at this conservative criterion, all eight DLT/PMT measures were significantly related to primary flight grades, however, none was significantly related to the pass/fail criterion. Regarding the background variables, the FAR stanine was significantly related to both pass/fail and flight grade. The Academic Qualifications Test was significantly related only to flight grade, and that relationship was not as strong as the FAR's r. The importance of attention to the form of the variable distributions is illustrated well with the background variable of previous flight hours. The most skewed of all the variables (cf. Tables 3 and 4), the raw score form of previous flight hours has an r of .12 with the original flight grade (see Table 6), but a log transform correlated .24 with the flight grade in z score form--representing a 100% increase in the value of r.

TABLE 9. Correlations of Log-transformed DLT/PMT Measures and Background Variables with Criteria.

Variable	Pass/Fail[a]	Flight grades Orig.[b]	Flight grades z score[c]
Dichotic Listening--log_{10} (Number Wrong + 1)			
Single DLT	.01	-.21	-.22*
Dual DLT1 (with PMT stick)	-.06	-.24	-.24*
Dual DLT2,3 (with PMT stick/rudder)	-.06	-.27	-.28*
Psychomotor Tasks--log_{10} Cumulative Pixel Errors			
Single PMT stick	-.10	-.33	-.34*
Dual PMT stick (with DLT1)	-.04	-.33	-.33*
Single PMT stick/rudder	-.07	-.39	-.39*
Dual PMT stick/rudder (with DLT2,3)	-.06	-.40	-.41*
Single PMT stick/rudder/throttle	-.11	-.30	-.26*
Background Variables			
Age	-.04	-.07	-.10
log_{10} previous flight hours	.05	.22	.24*
AQT (stanine)	.03	.15	.14
FAR (stanine)	.14*	.23	.27*
AQT (raw score)	.03	.15	.16*
FAR (raw score)	.13	.26	.29*

* Significant at .05 level.
[a] *n* = approximately 530 for all correlations with Pass/Fail except for PMT stick/rudder/throttle where *n* = 205.
[b] *n* = approximately 490 for all correlations with flight grades--Original except for PMT stick/rudder/throttle where *n* = 193. These correlations were not tested for significance.
[c] *n* = approximately 480 for all correlations with flight grades--*z* scores except for PMT stick/rudder/throttle where *n* = 185.

MULTIPLE REGRESSION ANALYSES

Several multiple regression analyses were conducted to characterize the joint relationship between the criteria and the various predictors. First, standard regressions were conducted in which all variables within a category (either DLT/PMT or background) were forced to enter as a block. This allowed the variability of each criterion to be partitioned into the components that could be accounted for by the performance-based tests, on the one hand, or the background variables, on the other. Secondly, stepwise procedures were used to see which predictors could be eliminated to achieve a more parsimonious model of the data.

In addition to the background variables summarized in the previous tables, three more nearly categorical variables were incorporated into the multiple regressions. They were: gender (1 = male, 2 = female), accession (1 = Naval Academy, -1 = AOC, 0 = otherwise), and educational major (1 = engineering or mathematics, 2 = general science, 3 = business, 4 = humanities/social science/psychology, 5 = physical education). Table 10 provides information on frequencies. These categories were used because of their monotonic relationship with the criteria, although the relationships were not generally strong. The exception was educational major, which correlated -.13 with pass/fail and -.14 with flight grades. Finally, with regard to the DLT/PMT variables, the single PMT stick/rudder/throttle measure was dropped because it was not as strongly related to flight grade as the other psychomotor variables, and keeping it in the analysis would have eliminated nearly two-thirds of the original sample.

TABLE 10. Frequency (and Percentage) of Gender, Accession, and Educational Major Categories in Original Sample.

Gender		Accession		Educational major	
Male	663 (98%)	Academy	141 (21%)	Eng/math	297 (44%)
Female	14 (2%)	AOC	315 (46%)	Gen sci	104 (15%)
		Other	221 (33%)	Business	155 (23%)
				Hum/SocSc	114 (17%)
				Phys Ed	3 (.4%)
				Missing	4 (.6%)

The multiple regression of the pass/fail criterion on the seven remaining (see Table 9) DLT/PMT variables was nonsignificant, $F(7, 501) = 0.97$, $p = .454$, $R = .116$. On the other hand, pass/fail was significantly predicted by a combination of the seven background variables of age, gender, accession, education, log previous flight hours, AQT stanine, and FAR stanine, $F(7, 501) = 2.90$, $p = .006$, $R = .197$. All 14 variables combined yielded $F(14, 494) = 1.92$, $p = .023$, $R = .227$. The addition of the background variables to the DLT/PMT variables resulted in a significant increase in R^2, $F(7, 494) = 2.84$, $p < .01$, but not vice versa. A graphical portrayal of the proportion of variance in pass/fail predicted by the sets of measures is presented in Fig. 3. The contribution of the sets of variables was essentially the same regardless of order of entry: DLT/PMT variables predicted 1.3% variance, background variables 3.9%.

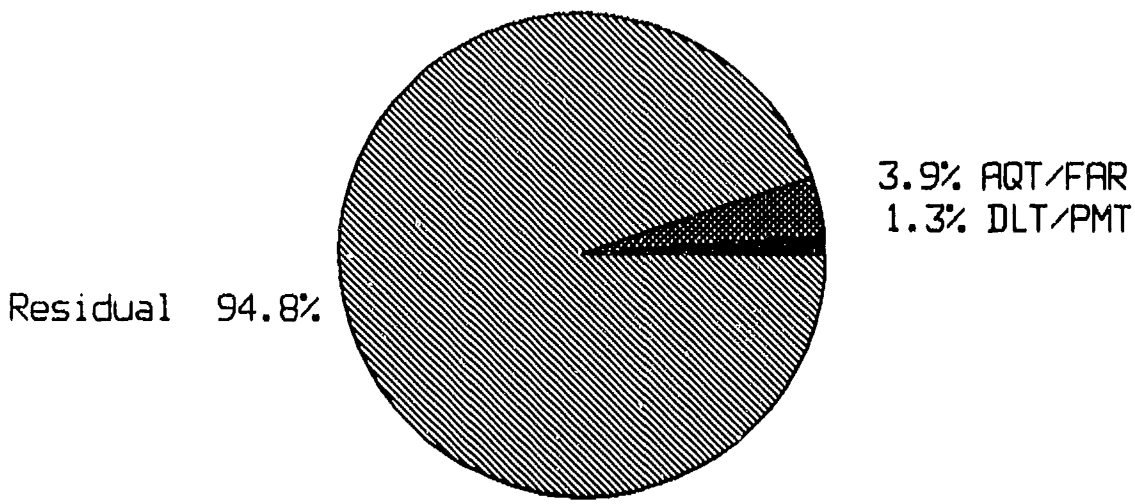


Figure 3. *Variance in pass/fail criterion accounted for by DLT/PMT variables and by AQT/FAR and demographic variables.*

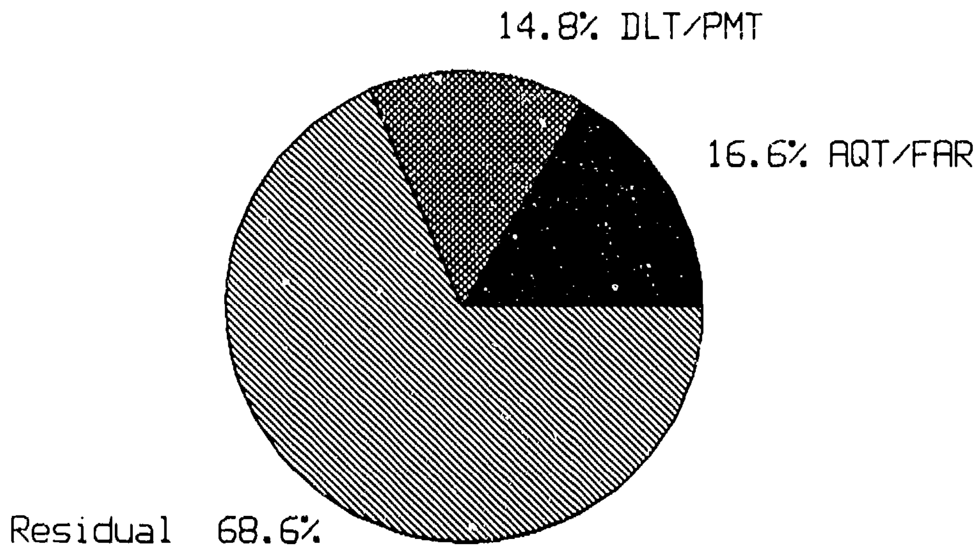
The forward stepwise regression analysis indicated that nearly the same multiple *R* could be achieved by a subset of seven variables, $F(7, 501) = 3.60$, $p < .0001$, $R = .219$. Standardized (beta) and unstandardized (*b*) weights are indicated in Table 11. The statistical criterion for entry of a variable into the equation was $t > 1$.

TABLE 11. Regression Weights for Predicting Pass/Fail.

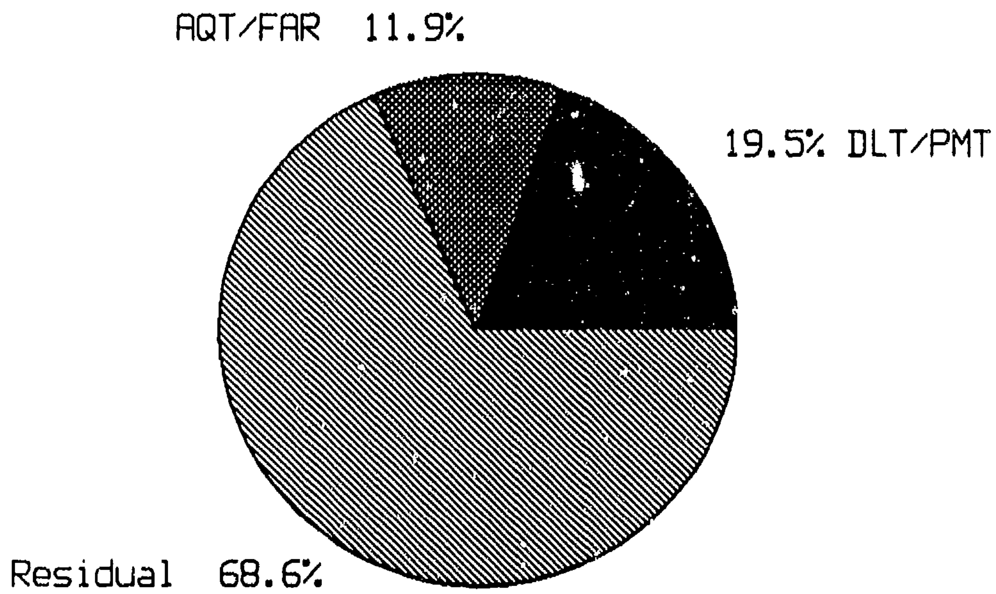
Variable	beta	b	t(501)	*p*
FAR stanine	.128	.022	2.72	.007
Education	-.085	-.021	-1.88	.061
Gender	.085	.158	1.87	.062
Accession	.076	.027	1.69	.091
log PMT stick	-.053	-.054	-1.10	.271
log Single DLT wrong	.102	.109	1.84	.067
log Dual DLT1 wrong	-.098	-.077	-1.73	.085
Intercept		.852		

Much larger multiple correlations were achieved with the flight grade criterion. When the DLT/PMT variables were used as predictors, $F(7, 456) = 15.78$, $p < .0001$, $R = .442$. Conversely, when the background variables were used as predictors, $F(7, 456) = 12.93$, $p < .0001$, $R = .407$. The combination of the two sets of predictors yielded $F(14, 449) = 14.69$, $p < .0001$, $R = .561$. Thus, as separate sets of predictors, the DLT/PMT measures accounted for 19.5% of the variance in flight grades, and the background measures accounted for 16.6%. The two sets together accounted for 31.4%. Graphical representations of these relationships are shown in Fig. 4. This indicates that the part of flight grades that DLT/PMT measures can predict is almost entirely independent in this sample from that predictable by the background variables. In particular, the increase in R^2 resulting from adding the DLT/PMT measures to the AQT/FAR and demographic measures is over 85% of that which would result if they were entirely independent. This increase in R^2 is highly significant, $F(7, 449) = 13.87$, $p < .0001$, as is that which results from adding the AQT/FAR/demographic variables to the performance-based tests, $F(7, 449) = 11.15$, $p < .0001$.

A stepwise regression analysis was also performed for the flight grade criterion. Ten predictors entered the final equation. These included three psychomotor variables. The most complex of the three, the dual PMT stick/rudder, was the most heavily weighted. The complete set of weights is shown in Table 12.



(a) Increment in flight grade variance accounted for by adding DLT/PMT scores to AQT/FAR and demographic variables.



(b) Increment in flight grade variance accounted for by adding AQT/FAR and demographic variables to DLT/PMT scores.

Figure 4. *Partitioning of variance in multiple regressions to predict flight grades.*

TABLE 12. Regression Weights for Predicting Flight Grade.

Variable	beta	b	t(453)	p
log Dual PMT stick/rudder	-.198	-.029	-3.30	.001
FAR raw score	.171	.001	4.00	.001
log Previous Flight Hours	.218	.011	5.20	.001
Accession	.108	.005	2.52	.012
log Single PMT stick/rudder	-.112	-.019	-1.69	.092
AQT raw score	.103	.001	2.47	.014
Gender	.081	.018	1.97	.049
log Single PMT stick	-.114	-.015	-2.11	.036
Age	-.072	-.002	-1.70	.089
log Single DLT wrong	-.070	-.009	-1.66	.098
Intercept		3.278		

PRACTICAL BENEFITS

Research in psychometrics and applied psychology over the past 50 years, which is reviewed by Schmidt et al. (17), has made clear that the practical benefit of implementing a valid selection test is proportional to the validity coefficient. The validity coefficient is the correlation between the test, or the prediction derived from a combination of subtests, on the one hand, and a criterion variable, on the other. Schmidt et al. indicate that this is true whether one considers the economic value or utility of the increased productivity achieved by selecting superior personnel, or whether one considers the economic value or utility of decreased attrition resulting from the selection procedure. The estimation of economic benefits of increased productivity would require one first to solve the problem of determining the value of, for example, having a trained pilot at a given level of skill. The benefits of using valid tests for reducing attrition is more straightforward to determine, at least in terms of estimating the proportion of individuals who will attrite.

In fact, approaching the issue mathematically, by making assumptions about the form of the predictor and criterion distributions, tables have been developed for translating validity coefficients into the proportion of applicants who would succeed if the test were used in selection (18). As a rough guide, at least in one special case the validity coefficient will equal the difference between the success rate for those the test indicates should have been accepted and the success rate for those the test indicates should have been rejected (cf. 19).

Because we have data on actual success rates, we can approach the problem empirically rather than mathematically. Thus, we can address the question of the practical benefit of the current tests by using the proportion of the attrites we could have identified as a basis for projecting to future reductions in attrition that would result if these selection tests were implemented.

The regression weights in Table 11 were used to compute a predicted pass/fail score[1] for each of the over 500 student naval aviators who completed or attrited from primary flight training in the current study. A helpful way of visualizing the results of this procedure is by displaying the two distributions of predictions, one for those who eventually passed and one for those who eventually failed, side by side in a single figure (see Fig. 5). This is analogous to the overlapping distributions of signal and noise used in signal detection theory (e.g. 20, Ch. 6). The displayed distributions indicate a reasonable amount of discriminability. In fact, as indicated in Table 13, the group means differ by close to one within-group standard deviation.

TABLE 13. Statistics on Predicted Pass/Fail Score for Pass and Fail Groups.

Actual Outcome	Mean	*SD*	*n*
Pass	.915	.061	478
Fail	.864	.060	46

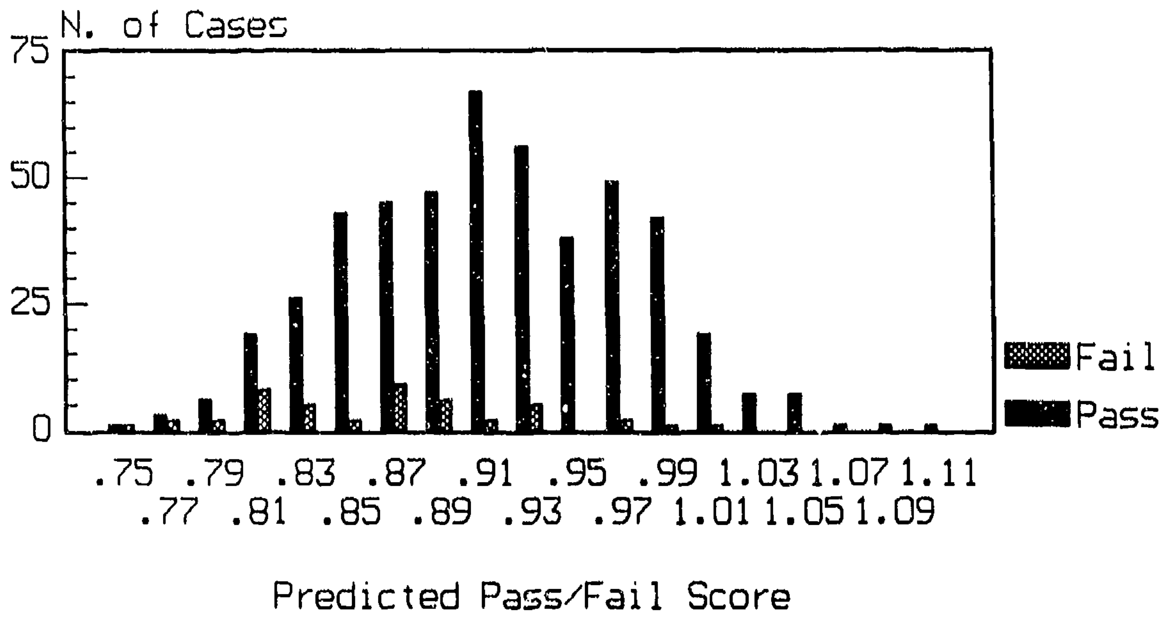


Figure 5. *Distributions of predicted pass/fail scores for 478 SNAs who passed and for 46 SNAs who attrited from primary flight training.*

[1] These predicted pass/fail scores may be thought of roughly as predicted probability of success. Because the predicted scores are simply the optimal linear function of the predictors, these predicted values on occasion can, however, unlike probabilities, be greater than 1.

A classification matrix, however, may provide the most useful summary of the results. In discriminant analysis terminology, this indicates the frequency with which cases in each group can be correctly identified. More to the point in a personnel selection situation, it indicates the success and attrition rates in identifiable subgroups that may be used to project the possible gains from actually implementing these tests for selection.

The classifications are a consequence of a decision to place a cutoff for selection into or rejection from the program at a given point. In terms of Fig. 5, this means deciding on a particular predicted pass/fail score below which candidates will be rejected. If the costs of misclassifications and base rate of success for the tested population are known, then decision theory procedures may be used to determine an optimal cutoff point (21). In the absence of this information, various plausible cutoff points can be tried.

One reasonable approach is to base the location of the cutoff point on the values implicit in the Navy's current operating procedures. For example, in recent years a "3,5" cutoff for AQT/FAR stanines has been a commonly referred to minimum standard for selection. If one were to derive predicted probability of success distributions like those in Fig. 5 but using only AQT/FAR stanines as predictors, a "3,5" cutoff would be equivalent to using a .85 predicted pass/fail score as the cutoff. Adopting this cutoff in the current sample and deriving predicted pass/fail scores only from the AQT/FAR stanines yields the classification matrix of Table 14 and provides a baseline against which to judge the benefits of a more complex decision rule. Thus, the actual overall attrition rate of 8.85% would have been 8.43% if everyone below the "3,5" cutoff on the AQT/FAR had actually been excluded (or equivalently if all those with a predicted pass/fail score below .85 were excluded, where predictions were based only on the AQT/FAR).

In contrast, using the predicted pass/fail score derived from the optimal weighting of DLT/PMT and background variables (see Table 11 and Fig. 5), yields the classification matrix of Table 15 when the .85 pass/fail-score cutoff is adopted. Thus, using the .85 cutoff based on an optimal weighting of predictors here would have been expected to reduce the attrition rate to 6.34%, or to approximately three-fourths of its current value.

TABLE 14. Classification Matrix Resulting from Using "3,5" Cutoff on AQT/FAR in Current Sample.

Decision	Predicted pass/fail score	Outcome: Pass	Outcome: Fail	Total
Accept	> .85	456 (91.57%)	42 (8.43%)	498
Reject	< .85	28 (84.85%)	5 (15.15%)	33
	Total	484 (91.15%)	47 (8.85%)	531

TABLE 15. Classification Matrix Resulting from Using Cutoff of .85 Pass/Fail Score Based on Optimal Combination of DLT/PMT and Background Variables.

Decision	Predicted pass/fail score	Outcome Pass	Fail	Total
Accept	> .85	399 (93.66%)	27 (6.34%)	426
Reject	< .85	79 (80.61%)	19 (19.39%)	98
	Total	478 (91.22%)	46 (8.78%)	524

CONCLUSIONS AND RECOMMENDATIONS

A large-scale validation of the dichotic listening and psychomotor tasks was carried out and was supportive of the value of these tests. Some statistical problems in both the predictors and the criteria were identified and addressed. Logarithmic transformations of the predictors largely solved problems of skewness and nonlinearity; between-squadron differences in flight grades required z-score transformations. The very high pass rate and variable factors influencing individual attritions necessarily limit the predictability of this criterion. Nonetheless, multiple regression results indicated that both pass/fail and primary flight grade could be significantly predicted by combinations of DLT/PMT and background variables. Psychomotor, dichotic listening, paper-and-pencil tests, and demographic variables all entered the final regression solutions for both pass/fail (Table 11) and flight grade (Table 12).

Finally, distributions of predicted probability of success and classification matrices were used to provide indications of the practical benefits that could be derived from using a regression-based decision rule for selection. The particular classification matrix used to illustrate the point was based on cutoffs implicit in the Navy's selection system. The benefit of reducing attrition to three quarters of its current value could be purchased at the cost of rejecting a relatively small proportion of candidates now allowed to enter primary flight training. The cutoff illustrated herein of a .85 probability of success would have eliminated 18.7% (98 of 524) of the current sample. This is in the range of rejection rates used by Kantor and Carretta (5) to illustrate the value of a proposed screening system for the U.S. Air Force and by Gopher (22) to illustrate the value of a selection procedure for the Israeli Air Force.

Clearly, alternate cutoff scores could be used. The high base rate of passes, at least among the samples tested at NAMRL, argues for a lenient criterion. However, base rate considerations are at least partially if not entirely offset by the greater cost associated with "false alarms", as compared with "misses." In other words, the cost to the Navy of partially training an individual who attrites is greater than the cost of testing and rejecting an individual who could make it through training.

In any case, the validity of the combination of measures analyzed is sufficiently high to result in tangible benefits from using the tests as selection devices. The observed multiple *R* of .219 for predicting pass/fail must surely be regarded as a lower-bound of the validity of the combination of measures used here. Although some shrinkage of *R* for the Table 11 prediction equation might be expected in a cross-validation in this same restricted population (the adjusted *R* estimating the population value is .186), this would be offset by the increase in *R* expected from using the tests in a less restricted population (13).

The current validation of the psychomotor component of aptitude for flying is simply the most recent in over 40 years of such demonstrations. The ready availability and high reliability of microcomputer-controlled testing now make such assessments eminently practical. The Air Force has recently committed to having a battery of tests including a psychomotor component operational for selection within 2 years (e.g., 23). The Navy would likely benefit from a similar commitment.

REFERENCES

1. Fleishman, E.A., "Psychomotor Selection Tests: Research and Application in the United States Air Force." *Personnel Psychology*, Vol. 9, pp. 449-467, 1956.

2. Cronbach, L.J., *Essentials of Psychological Testing*, Harper and Row, New York, NY, 1949, p. 221.

3. Petho, F.C., Collyer, P.D., and Sanders, C.M., "Psychomotor Test Performance in Primary Flight Training." *Proceedings of the Annual Scientific Meeting of the Aerospace Medical Association*, San Antonio, TX, 1981, pp. 278-279.

4. Carretta, T.R., "USAF Pilot Selection and Classification Systems." *Aviation, Space and Environmental Medicine*, Vol. 60, pp. 46-49, 1989.

5. Kantor, J.E. and Carretta, T.R., "Aircrew Selection Systems." *Aviation, Space, and Environmental Medicine*, Vol. 59, pp. A32-A38, 1988.

6. Bory, A. and Goodman, L.S., "Validation of a Performance Based Pilot Selection System." *Proceedings of the Annual Scientific Meeting of the Aerospace Medical Association*, Houston, TX, 1983, pp. 185-186.

7. Griffin, G.R. and McBride, D.K., *Multitask Performance: Predicting Success in Naval Aviation Primary Flight Training*, NAMRL 1316, Naval Aerospace Medical Research Laboratory, Pensacola, FL, March 1986.

8. Griffin, G.R., *Development and Evaluation of an Automated Series of Single- and Multiple-Dichotic Listening and Psychomotor Tasks*, NAMRL-1336, Naval Aerospace Medical Research Laboratory, Pensacola, FL, December 1987.

9. Gopher, D. and Kahneman, D. "Individual Differences in Attention and the Prediction of Flight Criteria." *Perceptual and Motor Skills*, Vol. 33, pp. 1335-1342, 1971.

10. Griffin, G.R. Mosko, J.D., Harris, S.D., Jones, T.N., North, R.A., and Owens, J.M, *Psychometric Properties of Dichotic Listening and IMPACT Tests: Intercorrelations, Reliabilities, and Relationship to Naval Aviation Selection Tests*, unpublished report, Naval Aerospace Medical Research Laboratory, Pensacola, FL, 1979.

11. Griffin, G.R. and Mosko, J.D., *Preliminary Evaluation of Two Dichotic Listening Tasks as Predictors of Performance in Naval Aviation Undergraduate Pilot Training*, NAMRL-1287, Naval Aerospace Medical Research Laboratory, Pensacola, FL, July 1982.

12. Fisher. R.A., *Statistical Methods for Research Workers*, Oliver & Boyd, Edinburgh, 1932, p. 74.

13. Guilford, J.P., *Fundamental Statistics in Psychology and Education*, McGraw-Hill, New York, 1956, pp. 148, 322.

14. Nunnally, J.C., *Psychometric Theory*, McGraw-Hill, New York, 1967, p. 133.

15. Doll, R.E., "Estimating the "True Validity" of the Naval Aviation Selection Tests." *Proceedings of the Annual Scientific Meeting of the Aerospace Medical Association*, pp. 24-25, 1977.

16. Maxwell, S.E. and Delaney, H.D., *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Wadsworth, Belmont, CA, 1990.

17. Schmidt, F.L., Hunter, J.E., McKenzie, R.C., and Muldrow, T.W., "Impact of Valid Selection Procedures on Work-force Productivity." *Journal of Applied Psychology*, Vol. 64, pp. 609-626, 1979.

18. Taylor, H.C. and Russell, J.T. "The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection," *Journal of Applied Psychology*, Vol. 23, pp. 565-578, 1939.

19. Rosenthal, R. and Rubin, D.B., "A Simple General Purpose Display of Magnitude of Experimental Effect." *Journal of Educational Psychology*, Vol. 97, pp. 527-529, 1985.

20. Coombs, C.H., Dawes, R.M., and Tversky, A, *Mathematical Psychology: An Elementary Introduction*, Prentice-Hall, Englewood Cliffs, NJ, 1970.

21. Darlington, R.B. and Stauffer, G.F., "A Method for Choosing a Cutting Point on a Test." *Journal of Applied Psychology*, Vol. 50, pp. 229-231, 1966.

22. Gopher, D., "A Selective Attention Test as a Predictor of Success in Flight Training." *Human Factors*, Vol. 24, pp. 173-183, 1982.

23. Naval Aerospace Medical Research Laboratory Memorandum by D.L. Dolgin summarizing Broad Area Review at Randolph AFB Air Training Command 18-20 July 1989, 24 July 1989.