2

NAMRL- 1331

INSTRUCTOR PILOT EVALUATIONS OF KEY NAVAL PRIMARY FLIGHT TkAINING CRITERIA

D. L. Dolgin, G. D. Gibb, T. Nontasak, and W. R. Helm

June 1987

NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
PENSACOLA FLORIDA

The opinions and interpretations contained herein are those of the authors and do not necessarily represent the views, policies, or endorsement of the Department of the Navy or any other government agency.

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.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

AD-A183 773

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved or public release;
distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
NAMRL-1331

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Naval Aerospace Medical
Research Laboratory

6b. OFFICE SYMBOL *(If applicable)*
03

7a. NAME OF MONITORING ORGANIZATION
Naval Medical Research
and Development Command

6c. ADDRESS *(City, State, and ZIP Code)*
Naval Air Station,
Pensacola, FL 32508-5700

7b. ADDRESS *(City, State, and ZIP Code)*
Naval Medical Command
National Capital Region
Bethesda, MD 20814-5044

8a. NAME OF FUNDING/SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS *(City, State, and ZIP Code)*

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
63706N	M0096001	1051	

11. TITLE *(Include Security Classification)*
INSTRUCTOR PILOT EVALUATIONS OF KEY NAVAL PRIMARY FLIGHT TRAINING CRITERIA

12. PERSONAL AUTHOR(S)
D. L. Dolgin, G. D. Gibb, T. Nontasak*, and W. R. Helm

13a. TYPE OF REPORT
Interim

13b. TIME COVERED
FROM 1986 TO 1987

14. DATE OF REPORT *(Year, Month, Day)*
June 1987

15. PAGE COUNT
11

16. SUPPLEMENTARY NOTATION
*American Society for Engineering Education Postdoctoral Fellow

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Selection, Criterion, Grades, Primary Flight Training

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The purpose of this study was to identify the graded training items that instructor pilots use to discriminate good from poor performance among students during primary flight training. We expected that the identification of those training items that accounted for most of the variance in performance measurement would provide more effective criteria for validation of pilot selection variables. Analysis of instructor pilot's survey ratings identified 24 of the 153 training items in the flight training syllabus as being most frequently graded above or below average. Only nine of the graded items however, had ratings significantly larger than their respective training stage mean ratings with some inter-rater agreement. Comparable findings were obtained with a preliminary analysis of assigned student grades. Results indicated that it was not possible to identify a strong cluster of training items that accounted for the majority of variance in grading in each training stage. It was apparent from the findings that a high percentage of training items within each stage accounted for at least some variance.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
X☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
J. O. Houghton, CAPT MC USN

22b. TELEPHONE *(Include Area Code)*
(904) 452-2286

22c. OFFICE SYMBOL
00

83 APR edition may be used until exhausted.
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

★U.S. Government Printing Office: 1985—507-047

UNCLASSIFIED

SUMMARY PAGE

THE PROBLEM

The purpose of this study was to identify the graded training items that instructor pilots use to discriminate student performance during primary flight training. Identification of specific graded items that account for most of the variance in performance measurement would provide effective criteria for validation of pilot selection variables.

FINDINGS

Analysis of instructor pilot's ratings of questionnaire items developed directly from flight training grading forms identified 24 of the 153 graded training items in the flight training syllabus that were said to be most frequently graded above or below average. Only nine of the graded items, however, had means significantly larger than their respective training stage means with moderate inter-rater agreement. The instructor pilot questionnaire findings were generally supported by results obtained from grades analyzed from aviation training jackets. Based on the present effort it was not possible to identify a strong cluster of training items that were typically used to discriminate good versus poor student performance. Although several items could be identified that were always graded as average (reflecting no discrimination between students), most training items reflected at least some degree of deviation from average on the grading scale. Using only the most discriminative training items as a criterion would omit many items that account for at least some variation in performance, and would probably introduce a bias in the evaluation of students' performance. In order to develop an alternative criterion to the overall cumulative flight grade, several training items would need to have been identified that accounted for the majority of differences in student pilot performance.

RECOMMENDATIONS

1. Graded items that were identified as most frequently above or below average should be evaluated further for use in augmenting the existing pass/fail and cumulative flight grade criterion data base.

2. Additional research should be directed toward evaluating the reliability of instructor assigned grades.

Acknowledgments

The authors wish to express their appreciation for the support and enthusiasm of the instructors and staff of Training Squadrons TWO, THREE, and SIX, Naval Air Station, Whiting Field, Florida. The research reported in this paper was completed under the Naval Medical Research and Development Command work unit 63706N.M0096.001-1056.

INTRODUCTION

Since World War II, considerable research effort has been directed toward predicting success in military aviation training. Currently, the selection of candidates for naval aviation training begins with the meeting of certain minimum physical and medical standards. This selection process also includes scores on four pencil-and-paper tests: the Aviation Qualification Test (AQT), Mechanical Comprehension Test (MCT), Spatial Apperception Test (SAT), and the Biographical Inventory (BI). The latter three tests constitute the flight aptitude rating (FAR). These test scores and physical standards determine which applicants are selected for training, and provide a probability estimate for a given applicant concerning completion of the aviation training program.

In spite of technological advances in selection techniques, little has been achieved concerning performance measurement in aviation training environments (1). Without valid, reliable, and discriminative measures of aviation training performance, it is unlikely that selection systems can be adequately validated. A major shortcoming of previous efforts to increase the predictive validity of aviation test selection systems can be directly related to methodological difficulties. More specifically, test measurement variables have been related to global criterion performance measures in training such as pass/fail or composite flight grades (2). These criteria, although very useful, are comprised of a number of undesirable psychometric properties and may obscure the components of skilled performance or behavioral attributes associated with new selection test measures. It is reasonable to assume that a given test measure would be highly related to a critical dimension of performance during some phase of training, but that the insensitivity or impracticality of the criterion variables used may yield low correlations and a consequent dismissal of the test's utility. This situation can be expected not only for the more traditional paper-and-pencil measures, but also for performance-based test measures.

The purpose of this study was to identify graded items that instructor pilots use in their discrimination of performance among students during primary flight training. The instructor pilot's function is to teach and evaluate student pilot flying performance. Thus, critical information regarding a student's performance is embedded in these instructor-assigned grades. Due to grading scales that have a limited range and the fact that all grades have equal weight in a composite flight grade, the range of student cumulative flight grades tends to be very restricted. An instructor pilot survey was designed to identify items that were graded differently (accounted for more variance) most frequently. Such items may result in new criterion measures with increased predictive validity. Several studies (3,4,5) have found that instructor ratings increased the reliable prediction of training success beyond that obtained using standard selection variables (AQT/FAR) and academic grades available prior to primary flight training. Using instructor identified variables to develop a criterion may serve to facilitate selection test battery validation efforts (6).

METHOD

SUBJECTS

One hundred thirty instructor pilots from the three fixed wing squadrons of Training Wing FIVE based at Naval Air Station, Whiting Field, Florida (NASWF), were administered a questionnaire. The group consisted of lieutenants junior grade (12%), lieutenants (80%), lieutenant commanders (7%), and commanders (1%).

The amount of experience since designation as an aviator ranged from 9 to 209 months (M = 66.3, SD = 37.8). The amount of experience as an instructor pilot at NASWF ranged from 1 to 84 months (M = 14.6, SD = 11.7).

PROCEDURES

A questionnaire was developed directly from NASWF flight training grading forms used by instructor pilots to evaluate all flights and maneuvers within the primary flight training syllabus. The questionnaire consisted of one form for each of the six stages of primary flight training: Familiarization (FAM), Basic Instruments (BI), Radio Instruments (RI), Precision Aerobatics (PA), Formation (FORM), and Night Familiarization (NFAM). The forms in the questionnaire were the flight grading forms, which were organized by training stages. These grading forms listed the maneuvers and procedures (training items) performed during all flights. The training items within each stage were rated by the instructors according to the extent that the grade assigned for the item was not graded average (A). That is, these grades were more likely to be either unsatisfactory (U), below average (BA), or above average (AA) on an interval scale. Scale values ranged from 1 (little or no extent) through 5 (very strong extent). Training items that were not typically graded or were not applicable were rated as "0." Instructors evaluated all training items on the basis of personal experience in grading. Instructors were briefed on the anonymity of the questionnaire when it was administered. They were instructed to respond only to those sections of the questionnaire that pertained to the training stages that they were currently teaching. Verifying data were obtained from 76 aviation training jackets drawn at random from each of the 3 fixed-wing squadrons for those students who had completed at least the FAM stage of training. We recorded the combined frequency that each training item in the FAM stage was graded U, BA, or AA, and the frequency that each individual training item could be graded. The relative frequency that each training item was graded either U, BA, or AA was computed by dividing the former measure by the latter measure.

RESULTS

To assess the extent that an item had ratings other than average, the data were analyzed by comparing each item mean against its respective stage mean. Then, the items rated more than 1 SD above their mean were selected. The means and standard deviations of the rated items for each training stage are given in Table 1. These means were obtained by calculating the mean ratings for all items within each training stage and then calculating stage means as an average of the individual item means. Of 153 graded items, 24 item means exceeded their respective stage mean by 1 standard deviation.

TABLE 1. Means and Standard Deviations of the Extent That Items Are Not Graded as Average, by Stage.

Stage	Number of training items	M	SD	Rated items exceeding M by 1 SD
FAM	45	2.50	.83	8
BI	20	2.76	.69	3
PA	23	2.53	.60	2
RI	29	2.54	.62	4
FORM	24	2.46	.55	5
NFAM	12	2.80	.70	2
Total	153	-	-	24

Item ratings were then converted to standardized scores, normalized within raters, to remove scale attributes from the data. The standardized scores, when rank ordered, agree with the finding above that 23 of the initial 24 items identified were the highest rated. Bonferroni one-tailed t-tests (7) indicated that 12 of the 24 individual item means were significantly larger than their respective stage means at the .05 level (Table 2). However, in assessing the strength of agreement among instructors for the 24 identified items, only 9 of the 24 items had 60% or more of the responses rated '4' or '5.' This indicated some lack of agreement among the raters on how often an item was graded average (Table 2).

TABLE , Selected Frequency Distributions, Means and t-values of Identified Maneuver/Procedure Evaluations, by Stage.

	Percent item was evaluated			
	4	5	M	t-value
FAM				
Landing pattern/FF landing	28	46	4.1	7.92*
NF landing	33	35	4.0	6.78*
HAPL	36	28	3.8	6.06*
LAPL (C)	40	20	3.7	5.12*
PPEL	37	23	3.6	5.14*
Procedures	16	36	3.6	4.69*
Basic airwork	24	24	3.4	3.97*
Approach turn stall	41	11	3.4	3.57
BI				
Basic airwork	37	33	3.9	5.21*
S-1	34	19	3.5	3.57*
Approach pattern	32	18	3.5	3.37*
PA				
Precision landing	47	32	4.1	4.00*
PPEL	42	8	3.2	1.32
RI				
Basic airwork	28	16	3.4	2.89
RVFAC	35	17	3.4	2.81
Procedures	24	20	3.3	2.30
Headwork	21	21	3.2	1.87
FORM				
Error correction	21	21	3.4	2.38
Procedures	37	18	3.4	2.30
Breakup & rendezvous	29	9	3.2	1.81
Parade position	18	8	3.1	1.70
Basic airwork	18	18	3.1	1.57
NFAM				
Night landings (5 minimum)	25	43	4.0	3.47*
Night landing pattern	30	37	3.8	2.88

* Significantly larger than corresponding stage mean, Bonferroni one-tailed, $p < .05$.

Instructor pilot questionnaire responses were subjected to a split-sample reliability technique (8). For each training stage, the original sample was randomly divided into two subsamples. Item means were then compared between subsamples for each training stage using a Pearson product-moment coefficient of correlation. This procedure resulted in correlations that ranged from .74 to .98 ($p < .001$ or better) between instructor pilot sub-groups, indicating that the questionnaire ratings were replicable.

Instructor pilot questionnaire ratings were compared with actual item grades obtained from 76 aviation training jackets randomly selected from the 3 training squadrons at NASWF. Table 3 delineates the frequency that each training item was graded other than A, and the number of occurrences it was graded. Training items are listed in descending order with respect to the relative frequency of collectively being graded as either U, BA, or AA. The data listed in Table 3 are limited to FAM because that stage had the largest number of training items (45) and the greatest percentage of training items significantly above the appropriate stage as determined by instructor questionnaire results.

TABLE 3. Frequencies and Relative Frequencies of Observed FAM Training Items Graded as U, BA, or AA ($\underline{n}$ = 76).

FAM training item	Number of times item is graded in stage	Frequency of occurrence of U, BA, AA combined	Relative frequency of U, BA, AA combined
Approach turn stall (ATS)*	9	142	.21
Preflight	3	47	.21
HAPL*	8	106	.17
Basic airwork*	13	161	.16
FF landings*	11	127	.15
Landing pattern*	11	118	.14
PPEL*	7	73	.14
PPEL(P)	5	47	.12
LAPL(P)	5	43	.11
Level speed changes	4	33	.11
Headwork	13	103	.10
NF landings*	10	63	.08
Straight and level	2	12	.08
Use/effect of controls/gears/flap	1	6	.08
Procedures*	13	76	.08
LAPL(C)*	6	34	.08
Spin	6	31	.07
Course rules/comm/IFF	13	63	.06
Power-off stall	3	14	.06
Taxi	3	13	.06
OFO	8	29	.05
OFE	3	9	.04
CABT	2	6	.04
Takeoff/departure	13	38	.04
Turn pattern	12	33	.04
Checklists	3	7	.03
Waveoff	10	20	.03
HFE	5	8	.02
Basic trainsitions	3	4	.02
Emergency procedures	13	17	.02
OFD	3	3	.01
ICA	12	6	.007
Start	3	1	.004
EPL-use	5	1	.003
Runup	3	0	.00
Aircraft stability	1	0	.00
Visual scan patterns	1	0	.00
Emer. ext. of landing gear	1	0	.00
Slip	2	0	.00
Skidded turn stall	1	0	.00
Crosswing landings technique	1	0	.00
Fx while airborne	1	0	.00
R/C secured for solo flight	1	0	.00

*Critical items identified by instructors using questionnaire approach (see Table 2).

DISCUSSION

Twenty-four of the 153 possible training items in the flight training syllabus were identified as those items most frequently scored above or below average according to the self-report questionnaire. However, nine of these items indicated only moderate agreement among instructor pilots, although the results indicated that the questionnaire responses were replicable. Thus, only nine training items differed significantly from their respective training stages and also had moderate inter-rater agreement. Although these nine items may reflect a proportionately higher amount of grading variance, a substantial proportion of the variance would be lost if only those items were used as a performance criterion. The omission of those training items which accounted for the variance in performance criterion, even if they were not significantly different or lacked agreement or reliability on the part of instructor pilots, could introduce a criterion bias commonly referred to as criterion deficiency (9).

We expected that several training items would have emerged that could account for most of the variance in grading, and that there would be high agreement among instructor pilots across all items in general. The data indicated, however, that inter-rater agreement was only moderate and that a high percentage of training items account for at least some of the variance. An explanation for the lack of agreement among instructors, other than such agreement being nonexistent, might involve the design of the survey itself. The possibility exists that the instructor pilots interpreted the survey differently from what was intended. The instructions did not clearly direct the instructors to indicate the training elements used by them to differentiate good from poor performance, only to indicate the training elements normally graded other than average. The possibility exists that some instructors were evaluating the training elements as they were typically graded. It is also possible that instructor pilots use different training items in grading student flight performance. A survey approach to identify specific graded items that could account for considerable variance in the measurement of student flight performance was not adequate to augment the current dichotomous pass/attrite or cumulative flight grade criterion data base. An alternative approach would have been to analyze an extensive flight performance criteria data base in order to identify graded items having the greatest variability. To do such a comprehensive evaluation would require prohibitive resources and time because the United States Navy does not currently maintain an automated flight training data base of all the grades that a student receives during flight training. This approach, although labor intensive, would rely completely on the actual grading practices of instructors and would remove the possibility of survey ambiguity. We attempted this approach with an analysis of the grading trends within the FAM stage hoping to verify the questionnaire findings. These results supported the instructor questionnaire findings in that of the 16 most discriminatory items from the jackets, 8 were identified from the questionnaire. As previously determined, several training items within the FAM stage deviated from the average grade proportionately more than the remaining items. Most of the training items that had higher relative frequencies for collectively being graded as U, BA, or AA were similar to the instructor-rated training items used to discriminate student flight training performance. Of the remaining training items, a high percentage accounted for at least some of the variance. These results indicated considerable differences among instructors in their evaluation and grading of individual training items.

CONCLUSIONS

Based on the present effort, it was not possible to identify a strong cluster of training items that were typically used to differentiate good from poor student performance. Although several items could be identified that were always graded as average (reflecting no differentiation between students), most training items reflected at least some degree of deviation from average on the grading scale. Using only the most discriminative training items as a criterion would omit many items that account for at least some variation in performance, and would probably introduce a bias in the evaluation of students' performance. In order to develop an alternative criterion to the overall cumulative flight grade, several training items would need to have been identified that accounted for the majority of differences in student pilot performance.

REFERENCES

1. Lane, N.E, Issues in Performance Measurement for Military Aviation with Applications to Air Combat Maneuvering, NTSC TR86-0008, Essex Corporation, Orlando, FL, 1986.

2. North, R.A. and Griffin, G.R., Aviator Selection 1919-1977, NAMRL SR77-2, Naval Aerospace Medical Research Laboratory, Pensacola, FL, October 1977.

3. Waag, W.L., Shannon, R.H., and Ambler, R.K., The Use of Confidential Instructor Ratings for the Prediction of Success in Naval Undergraduate Pilot Training, NAMRL 1175, Naval Aerospace Medical Research Laboratory, Pensacola, FL, February 1973.

4. Berkshire, J.R., Some Inference from Military Training Research, NAMRL TM63-9, Naval Aerospace Medical Research Laboratory, Pensacola, FL, 1963.

5. Shannon, R.H. and Wagg, W.L., The Prediction of Pilot Performance in the F-4 Aircraft, NAMRL 1186, Naval Aerospace Medical Research Laboratory, Pensacola, FL, July 1973.

6. Damos, D. and Gibb, G.D., Development of a Computer-based Naval Aviation Selection Test Battery, NAMRL 1319, Naval Aerospace Medical Research Laboratory, Pensacola, FL, August 1986.

7. Miller, R.G., Jr., Simultaneous Statistical Inference, McGraw-Hill Co., New York, 1966.

8. Anastasi, A., Psychological Testing, 4th ed, Macmillan Publishing Co., New York, 1976.

9. Brogden, H.E. and Taylor, E.K., "The Theory and Classification of Criterion Bias." Educational and Psychological Measurement, Vol. 10, pp. 159-186, 1950.

*U.S. GOVERNMENT PRINTING OFFICE: 1987-730-013/65158 Region 4.