FAA-AM-83-9

RADAR TRAINING FACILITY INITIAL VALIDATION

James O. Boone

Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Ok 73125

MAY 1983

Document is available to the public through the
National Technical Information Service
Springfield, Virginia 22161

Prepared for
U.S. DEPARTMENT OF TRANSPORTATION
Federal Aviation Administration
Office of Aviation Medicine
Washington, D.C. 20591

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

Technical Report Documentation Page

1. Report No.
FAA-AM-83-9

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
RADAR TRAINING FACILITY INITIAL VALIDATION

5. Report Date
MAY 1983

6. Performing Organization Code

7. Author(s)
James O. Boone

8. Performing Organization Report No.

9. Performing Organization Name and Address
FAA Civil Aeromedical Institute
P.O. Box 25082
Oklahoma City, Oklahoma 73125

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

12. Sponsoring Agency Name and Address
Office of Aviation Medicine
Federal Aviation Administration
800 Independence Avenue, S.W.
Washington, D.C. 20591

13. Type of Report and Period Covered

14. Sponsoring Agency Code

15. Supplementary Notes

Work was performed under task AM-C-82-PSY-82.

16. Abstract
The Radar Training Facility (RTF), part of the Federal Aviation Administration Academy located at the Oklahoma City Mike Monroney Aeronautical Center, is designed to identify, as early as possible, air traffic control specialists who do not demonstrate sufficient potential to perform at radar tasks at the journeyman level. An extensive initial validation effort involving design evaluation, implementation evaluation, and formative evaluation was performed to determine if the system was adequately designed from an educational point of view, if the measures employed were reliable, if the program had a proper concurrent relationship with nonradar Academy measures, and if the difficulty level was appropriate. While the program was not pass/fail during the study, a score of less than 70 was used to calculate experimental pass/fail rates. Data from the initial validation effort indicated that the system was educationally sufficient, reliabilities were low to moderate, the RTF program had a proper concurrent relationship with nonradar measures, and the difficulty level was approximately correct. Information from data collected appeared to be asymptoting, and it was recommended that the program could begin pass/fail. The data for this study were collected prior to the air traffic control specialist's strike, and following the strike the Academy radar phase was sequenced out of the basic training curriculum to a later point (after the developmental ATCS had successfully checked out at the field facility in a nonradar position). Since the data were collected while the Academy radar phase immediately followed the Academy nonradar phase, application of these results (from basic trainees) to ATCS developmentals with several months of field experience should be done with this consideration in mind.

17. Key Words
Radar Training
Air Traffic Control
Training Validation Model

18. Distribution Statement
Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages

22. Price

Reproduction of completed page authorized

RADAR TRAINING FACILITY INITIAL VALIDATION

Radar Training Facility (RTF) Rationale.

Successful air traffic control specialists (ATCS's) who have made a transition from manual to automated air traffic control (ATC) appear to prefer the advantages in the automated environment. However, some prospective ATCS's do not perform successfully in radar ATC. Successful employment in the radar environment requires that a person possess certain aptitudes. It is in the interest of the Federal Aviation Administration (FAA) and the prospective ATCS's to determine as soon as possible if the prospective ATCS possesses the aptitude necessary to successfully operate in the radar ATC environment. The philosophy of the FAA in regard to this selection process is that the best way to measure aptitude is to place the individual in a radar simulation laboratory and perform a systematic, objective appraisal of the person's potential. To this end the FAA constructed a Radar Training Facility (RTF) at the FAA Academy in Oklahoma City, Oklahoma. The initial plan was that, during the training/screening process, the trainee would receive enough basic radar training to allow systematic evaluation of his or her performance.

The data for this study were collected prior to the August, 1981 ATCS strike. Prior to the strike, an "up or out" pass/fail policy was in effect which required that training failures be separated from the agency. Subsequent to the strike a new policy was implemented that established a career nonradar ATCS position. This new policy had an impact on the RTF in two primary ways. First, the Academy radar training phase, which had been scheduled to immediately follow the nonradar Academy phase in basic ATCS training, was moved to a post-Academy-graduation training phase after the developmental ATCS had successfully checked out at the field facility in a nonradar position. To become eligible for selection to a radar position under the new policy, the developmental had to return to the Academy and successfully complete the Academy radar phase at the RTF. The second effect of this policy change was to eliminate the required termination of developmentals who did not successfully complete the Academy radar phase; the failures are now returned to their nonradar field position. The data in this study were collected while the Academy radar phase immediately followed the Academy nonradar phase. Consequently, application of the results of this study on basic trainees to ATCS developmentals with several months of ATC experience in a later Academy radar phase should be done with this consideration in mind. For example, pass rates would be expected to rise significantly for the more highly trained developmentals.

Development and Construction of the Radar Training Facility.

In July 1976, engineering requirements were completed by the FAA for a radar training system. During that same month the FAA Administrator approved the procurement and construction of the RTF to be located at the FAA Academy in Oklahoma City. In October 1977, the FAA completed a program implementation plan that outlined the development and implementation of the

RTF. The contract for the development of the computer-driven simulator training system was awarded to Logicon, Tactical and Training System Division, San Diego, California, in January 1978. Groundbreaking for the construction of the new RTF was held on December 22, 1977. The new facility was built and accepted by the FAA in January 1980, and the training system developed by Logicon Corporation was accepted in April 1980.

Development of a Validation Strategy.

Through a series of communiques between the Director, Personnel and Training, APT-1 and the Director, Aeronautical Center, AAC-1, dating from October 1977 through March 1978, the validation strategy for the RTF evolved. Briefly, the strategy was (i) to form a validation committee with members from the Civil Aeromedical Institute (CAMI), the Academy, the Office of Personnel and Training (APT), and the Air Traffic Service (AAT); (ii) the committee was to visit and review the U.S. military and the Canadian civilian radar training facilities; and (iii) the committee information was to be used by CAMI to develop an evaluation model.

The visits and reviews conducted by the committee revealed three important concepts that were given to CAMI for development of the validation model: (i) There was no standardized and objective measure of radar trainee performance available that was superior to the methods employed by the FAA Academy in nonradar training. Consequently, the methods employed in Academy nonradar evaluation should be the beginning point for developing radar training performance measures. (ii) The lack of sufficient input by measurement specialists in designing the reviewed radar training configurations had resulted in costly changes to the system, after system development. Thus, the committee recommended that CAMI attend all program reviews and participate in the development of the training configuration. (iii) Performance at the pseudopilot position was found to interact with performance at the radar position, and this problem led most facilities to hire pseudopilot operators rather than have trainee confederates operate the position. CAMI used these inputs in developing the validation model.

The Validation Model.

Model Description. The generic model developed for RTF ATCS training program evaluation consists of four components: (i) design evaluation, (ii) implementation evaluation, (iii) formative evaluation, and (iv) summative evaluation.

Design Evaluation. The goal of program design evaluation is the proper development of the educational attributes of the system, curricula, and performance measurement procedures that make up the program implementation plan.

Implementation Evaluation. The implementation evaluation phase is designed to monitor program implementation and to insure and document that the program was implemented strictly according to the implementation plan. Any changes made to the design during implementation are carefully documented and the design is revised accordingly. The implementation evaluation stage insures that the stated process is operational, intact, and stable. This evaluation is generally accomplished by means of frequent status studies during the implementation stage. Data are collected on each aspect of the process and a determination made about the state of implementation. The status studies are generally made into a report for decision-makers with suggestions to improve or expedite implementation. Shortcomings in implementation are noted in each report.

Formative Evaluation. When the program is determined to be operational, intact, and sufficiently stable, formative and summative evaluations begin. Formative evaluation is the ongoing process of collecting data and statistics related to training criteria, to determine how well students are doing in training. The analyses are used to gauge the operational stability of the program and the quality of students coming into the program. It is also a method for monitoring compliance with Equal Employment Opportunity Commission (EEOC) guidelines (7).

Summative Evaluation. Summative evaluation is a continual assessment of the quality of the products of the program. While formative evaluation is summarized on an input-by-input basis and serves as an immediate feedback loop for ongoing program revisions if needed, summative evaluation occurs on a larger scale across a longer time span (e.g., on a yearly basis). Formative evaluation is concerned with internal program accuracy and stability, program reliability, and content and/or concurrent demonstrations of validity. (For example, are the measures reliable? Are the objectives well matched with curricula content? Do the pass/fail rates remain stable?) Summative evaluation, however, is a check on the quality of the output from the stabilized program. The summative evaluation is a test of predictive or criterion validity. It is a measure of the on the job success of those who pass the Academy training, and the relationship of how well the candidates performed in the Academy compared to how well they performed on the job. The so-called validity coefficient is the measure of this relationship. A more detailed description of the model is available elsewhere (1).

Scope of the Model in the Initial Validation Study.

The aim of the initial validation was to determine whether (i) the measures used to assess student performance are reliable, (ii) the measures used are concurrently valid, (iii) the program is operating at a stable level, (iv) the program is operating at the desired level of difficulty to screen those who do not show sufficient potential to perform radar air traffic control, and (v) to assess the initial fairness of the program with regard to sex and minority status. To accomplish these goals, design evaluation, implementation evaluation (prototype classes), and two loops of formative evaluation were performed. As pointed out in the model description, these three components are designed to measure internal reliability and validity. The ultimate evaluation of the program's effect comes with the summative evaluation where the on-the-job performance of those who successfully complete the program are tracked to determine if the program produces quality products and is predictive of on-the-job success. The initial validation also serves as baseline data for future assessment of changes made in the program. The remainder of this report covers the design, implementation, and formative phases of the initial validation and a general review of data across evaluation phases.

Design Evaluation.

Two Branches of the Design Evaluation Phase. There were two main branches in the design and development process. One branch involved system hardware and software design and the second branch involved curriculum, course, and performance design. The design evaluation for hardware and software development was accomplished by the author's serving as a consultant to the program director and participating in program reviews and configuration control. The design evaluation for curriculum, course, and performance development consisted of the author's serving as consultant to the Academy development staff. Two research studies were also conducted as a part of the design evaluation phase. The purpose of the first study was to determine error rates on the new pseudopilot position and the feasibility of employing disabled persons to operate the position. The purpose of the second study was to develop an over-the-shoulder radar grading process and to assess the utility of computer-derived measures in student assessment.

Summary of Findings and Recommendations. From an educational technology point of view, the developed system is adequately designed to provide (i) a sufficient amount of realism for basic radar training to the degree that radar aptitude can be assessed, and (ii) sufficient flexibility to provide a wide enough variety of radar scenarios. During system testing it was determined that (iii) some enhancements to the system could improve the educational quality of the system, and (iv) some system downtime can be expected during graded problems.

Based on the study of the pseudopilot position (2) it was determined that (i) the position can be operated with sufficient accuracy without a serious interaction with performance at the radar position, (ii) some of the keys on the keyboard could be better placed or better utilized; however, the keyboard is in general adequately configurated for accurate inputs, and (iii) handicapped persons can be trained to operate the position accurately. (Note: Since the study, 27 handicapped persons were hired as pseudopilots. The program has been very successful; it has attracted several awards, including the Department of Transportation (DOT) handicapped employee of the year award, presented to two of the employees.)

The purpose of CAMI's initial study of RTF grading and training procedures (3), was to solidify the design of laboratory problems and the assessment process. Products from the study included (i) a set of refined laboratory problems to be used as a starting point in prototype RTF training, (ii) an initial laboratory grading manual, and (iii) an initial laboratory grading worksheet and evaluation form, including the error categories for grading radar performance. Results of data analyses from the study showed that: (i) Error categories for radar assessment should consist of (a) system errors (15 points), (b) system deviations (10 points), (c) procedure and coordination errors (5 points), and (d) other errors (1/2 point). The differentiation between system error and system deviation which is not used in nonradar evaluation is due to the ability of an instructor and radar trainee to visibly determine the difference in the two errors. Radar laboratory evaluation forms reflect that difference. (ii) The predictive information of laboratory problems peaks at four problems; however, adding a fifth problem contributes significantly to the prediction of radar ability. An initial decision was made in the interest of time savings to use four laboratory problems in the prototype classes. (iii) The initial weights for the four problems were set at 15, 25, 30, and 30, respectively. (iv) The use of computer-derived automated measures in assessing student performance is a feasible procedure; however, more development, primarily software development, would be required. (v) The reliability of the laboratory problems is about .43 (the average intercorrelation among lab scores), a low-to-moderate level. This coefficient is consistent, however, with past studies in ATCS assessment listed in the Institute for Defense Analysis (IDA) (4) study.

The initial curriculum design and assessment design of radar follows closely the process used in the nonradar training phase. There are blocks of academic instruction, with block tests and a comprehensive phase test (CPT) weighted 2 percent and 8 percent, respectively, in the composite score, a laboratory phase weighted 65 percent in the composite; and a Controller Skills Test (CST) weighted 25 percent. Laboratory problems are scored over-the-shoulder, by an instructor, on forms tailored to radar assessment with the laboratory problem score comprising an average of the technical assessment and the instructor assessment. The basic difference between the radar and nonradar lab involves the radar screen, on which students visually see the aircraft and associated ATC information related to each aircraft. The separation standards in the radar phase allow aircraft to fly closer together, and this increases the number of aircraft that the student controls in the radar setting.

In general, the design evaluation phase resulted in data that indicated the system was ready for implementation. However, the data also suggested a cause for (i) minor concern over the reliability of the laboratory scores and (ii) some additional concern over "bugs" in the computer system that caused system crashes which could influence laboratory grading if too numerous.

Implementation Evaluation.

Overview of the Process. The purpose of the implementation evaluation was to conduct a shakedown on the radar training facility system so that any needed changes could be made prior to stabilizing the program and collecting formative data on approximately 250 trainees. The desired result of the implementation evaluation was a proper implementation of the program design.

The prototype implementation classes consisted of 81 students in terminal and 38 students in en route. The evaluation staff at CAMI and the Academy staff met frequently during the training, making small changes in the program as deemed necessary. At the end of the evaluation phase, various statistics were calculated to assess the need for any major changes in the assessment components.

Summary of the Findings and Recommendations.

The implementation evaluation revealed some areas where changes should be made prior to stabilizing the program for a long enough period of time to run a formative evaluation loop. Findings are reviewed separately for academic scores and laboratory scores.

Academic Scores. Item and test characteristic measures showed that the CPT and CST in both terminal and en route needed improvement. Reliabilities ranged from .463 to .626. In general, reliabilities were lower for en route than for terminal. Item statistics assisted in isolating troublesome items that were poorly written or miskeyed. Further, the en route CST was found to be highly duplicative of the en route CPT. Recommendations included (i) a general review of the items and distractors for correctness and clarity, (ii) a check for matching information conveyed in the lesson plans, and (iii) the en route staff being advised to revamp their CST to contain applied ATC problems to solve and making them less similar to the CPT academic test.

Laboratory Scores. Laboratory score intercorrelations indicated that unreliability existed in the scoring procedures. (i) A review of the laboratory grading manual and instructor training in laboratory scoring was recommended. The means, standard deviations, and experimental pass/fail rates showed that, in general, the program was too difficult to achieve the desired passing rate of 85-90 percent. Further, it was noted that terminal problems 1 and 2 were more difficult than problems 3 and 4. (ii) A thorough review of the difficulty level for each problem was recommended. Considering the laboratory reliabilities and general difficulty level, it was suggested that a fifth problem be added to the laboratory assessment process. The first two problems were to be weighted 10 and 15 points respectively, and the last three problems were to be weighted 25 points each. Computer failures during graded problems were also noted during this evaluation phase. The length of time on position, especially since errors are more concentrated toward the end of the problem, can affect laboratory grades. (iii) It was suggested that unless the problem runs at least 20 minutes (problems are planned to run 30 minutes) that it be repeated with another version of the problem. The 20-minute run time is not viewed as a sufficiently stringent criterion for operating the actual pass/fail program; however, for formative data collection, practical considerations dictated this less stringent criterion.

Formative Evaluation (Loop 1).

Formative Evaluation Loops. Two loops of the formative evaluation were conducted. The program was unchanged during each loop long enough to get a sufficient sample of ATCS students for the program statistics to stabilize (approximately 150-200 trainees in each option). The formative statistics are extensions of the statistics presented in the implementation evaluation phase. The following summary is from the first formative loop.

Summary of Findings and Recommendations. The formative evaluation data showed that the program was moving nearer to the desired state.

Reliabilities were still somewhat low, and differential pass rates by minority status continued to produce some concern. Computer outages during this evaluation phase occurred more frequently than in the previous evaluation phase. Power conditioning and communication clogs between the PDP 11/60 and PDP 11/34's were checked as sources of the problem. However, the problem was not solved. Tests for equality of the three parallel versions of each laboratory problem indicated a lack of equality for some versions. Recommendations included (i) reviewing test items using the process described previously in the implementation evaluation, (ii) eliminating emergency procedures from laboratory problems to make them less difficult (thereby removing a barrier to assessing core skills; at this stage of training simulated emergency procedures interfered with proper assessment of basic skills), and (iii) reviewing in detail the three versions of each laboratory problem and implementing changes where necessary to make the versions equal in difficulty.

Formative Evaluation (Loop 2).

The Impact of Changes from Formative Loop 1. While experimental pass/fail was calculated for the RTF classes, true pass/fail conditions were not actually operational, i.e. no one was failed and scores were used only for research purposes. The effects on RTF scores of using persons who had recently completed an arduous nonradar pass/fail program, and who knew their performance in the RTF would not affect their career, is unknown. This aspect of the data was assessed in more detail in the second formative loop. The aim was to determine if changes made from the first formative loop appeared to have a significant impact on the second formative loop data, or if the effects of improvements had become asymptotic due to a stable and moderate amount of error variance (possibly created by reduced motivation in the test subjects).

Summary of Findings and Recommendations. In general, laboratory means and pass rates were lower for terminal and higher for en route trainees in formative loop 2 than was found in the previous formative loop. Item analyses in loop 2 also fluctuated in a nonconsistent pattern when compared to loop 1. These fluctuations in data were probably due to a combination of lower motivation than would be expected if the program were pass/fail and the lack of high measurement reliability in laboratory scores. There remains a significant difference in pass rates between the majority group and minority groups.

Since very little information appeared to have been added by collecting data on a second formative loop, it was assumed that information from the data collected in a non-pass/fail mode had asymptoted, and no further data collection was planned prior to an actual pass/fail implementation.

Review of Data Across Evaluation Phases.

The evaluation phases have provided information covering a number of areas involving program validation. In this section the information across evaluation phases is assimilated and presented in more detail to form an overview on each of the following issues: (i) experimental pass/fail rates,

(ii) reliability, (iii) fairness, (iv) computer outages during graded problems, (v) data reduction and analysis, and (vi) concurrent validity measures.

Experimental Pass/Fail Rates. Data from the second formative loop showed an overall pass rate for en route and terminal of about 80 percent. Applying the criterion of an increase in pass rates of .5 to 1 standard deviation, based on past CAMI experience in moving nonradar practice laboratory problems to graded problems, a pass rate of approximately 86-88 percent can be expected if the program were made pass/fail operational.

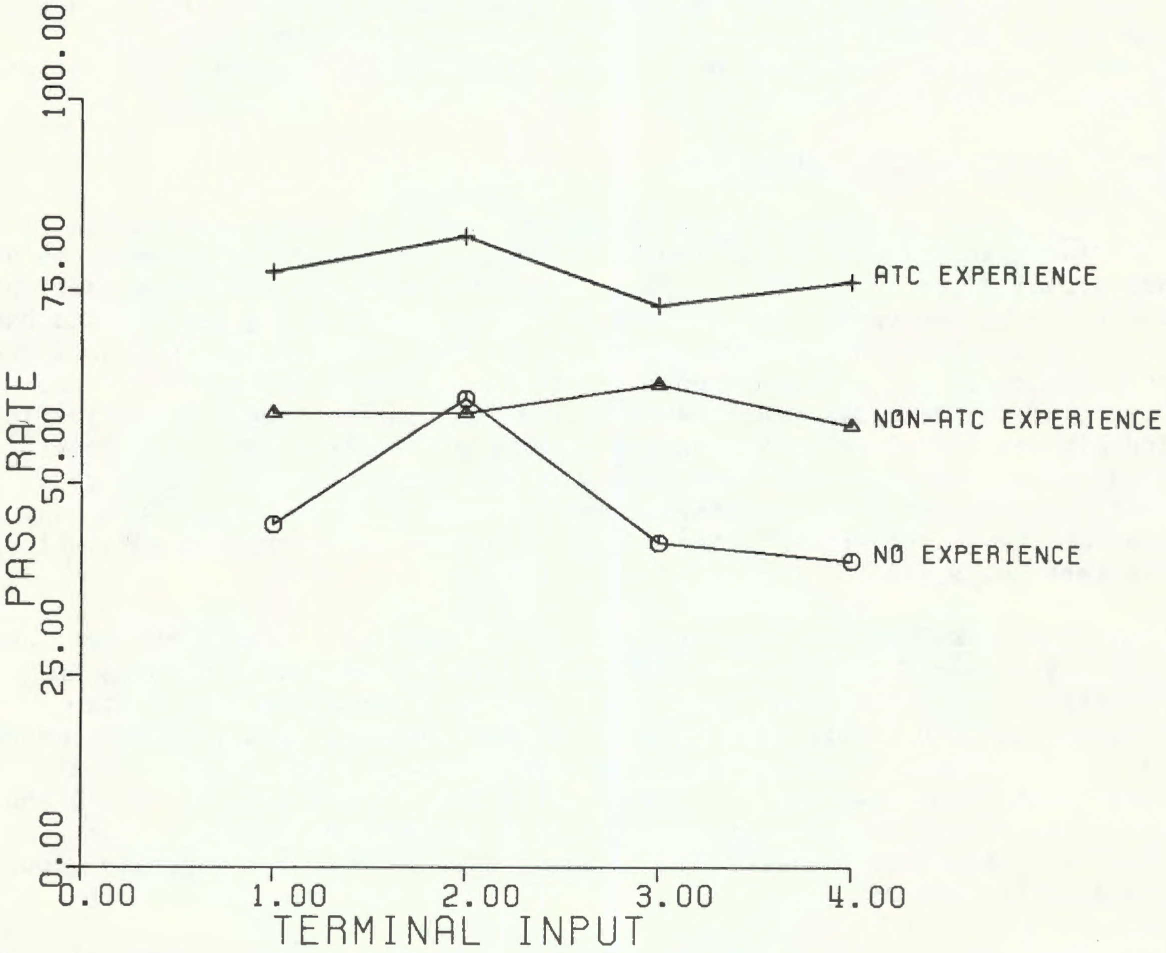

Figure 1. Pass Rates Across Evaluation Phases for Those with Prior ATC Experience (N=157), Related Aviation Experience (N=40), and No Experience (N=88) for the Terminal Option.

While pass/fail rates are not comparable between evaluation phases, due to the changes made in the program, pass rates within each phase can be compared for various subgroups across the phases. Figures 1 and 2 show the

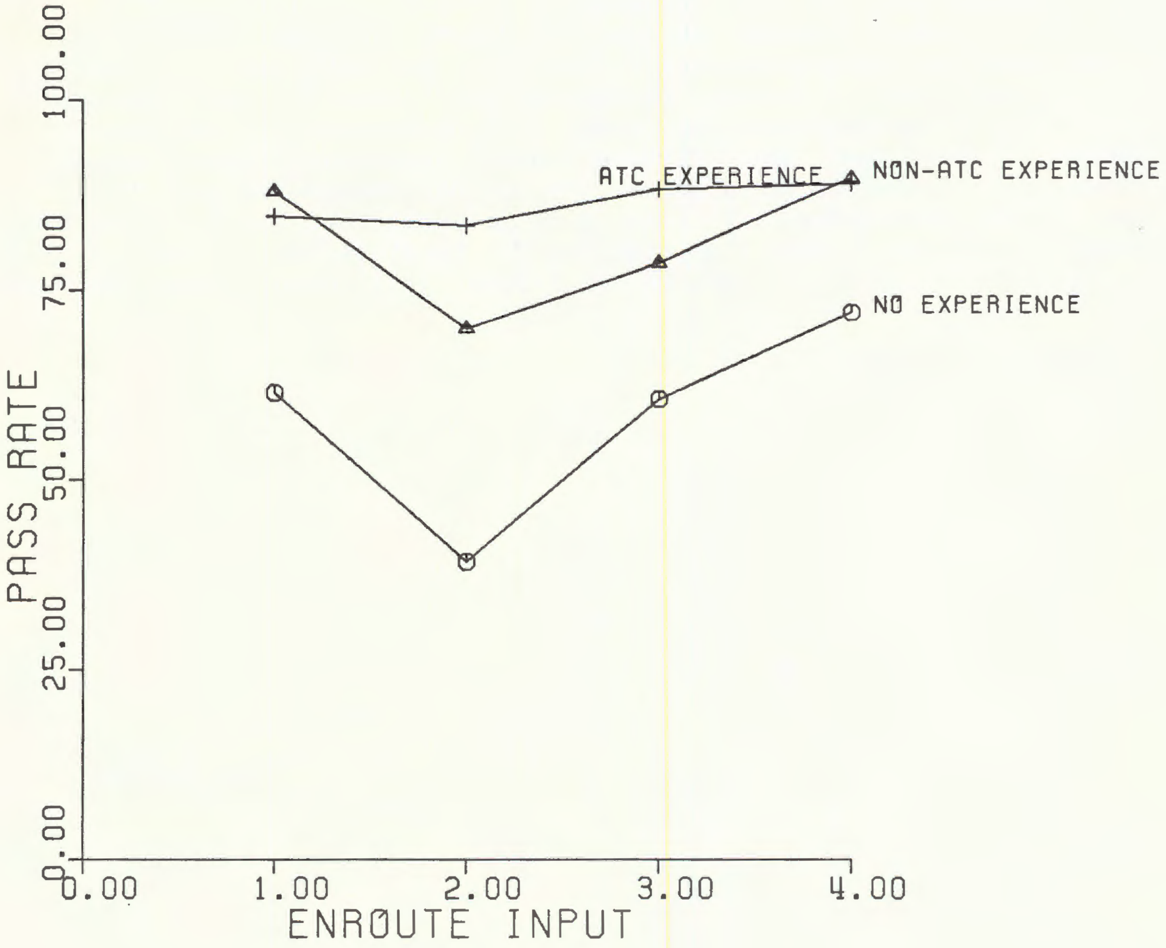

Figure 2. Pass Rates Across Evaluation Phases for Those with Prior ATC Experience (N=113), Related Aviation Experience (N=66), and No Experience (N=104) for the En Route Option.

pass rates for three subgroups across evaluation phases. The three subgroups consist of those with (i) no prior aviation-related experience, (ii) prior aviation-related experience, and (iii) prior ATC experience. If ATC radar skills were being appropriately measured, one would expect that the subgroup with prior military ATC experience (which is almost exclusively radar control) would have the highest pass rates, with prior aviation-related experience and no experience following in that order. The data in Figures 1 and 2 support this line of reasoning. These data are similar to nonradar pass rates; however, pass rates are increased more in radar by prior ATC experience than in nonradar, which, again, one would expect.

One confounding factor in radar pass/fail rates is an interaction effect between radar and nonradar pass/fail rates. Radar evaluation test subjects were recent Academy nonradar graduates. In some graduating nonradar inputs, there was a larger density of graduates who were near the pass/fail cut-score. The radar program tended to fail low-scoring graduates from nonradar. This interaction effect will be minimized with the new placement of radar training. Figures 3 and 4 graphically illustrate the nonradar and radar pass rates across the program evaluation periods. The terminal pass rates show an increase for the first interval for radar and nonradar, while the latter two intervals for terminal have an inverse interactive relationship; when nonradar pass rates go up, radar goes down.

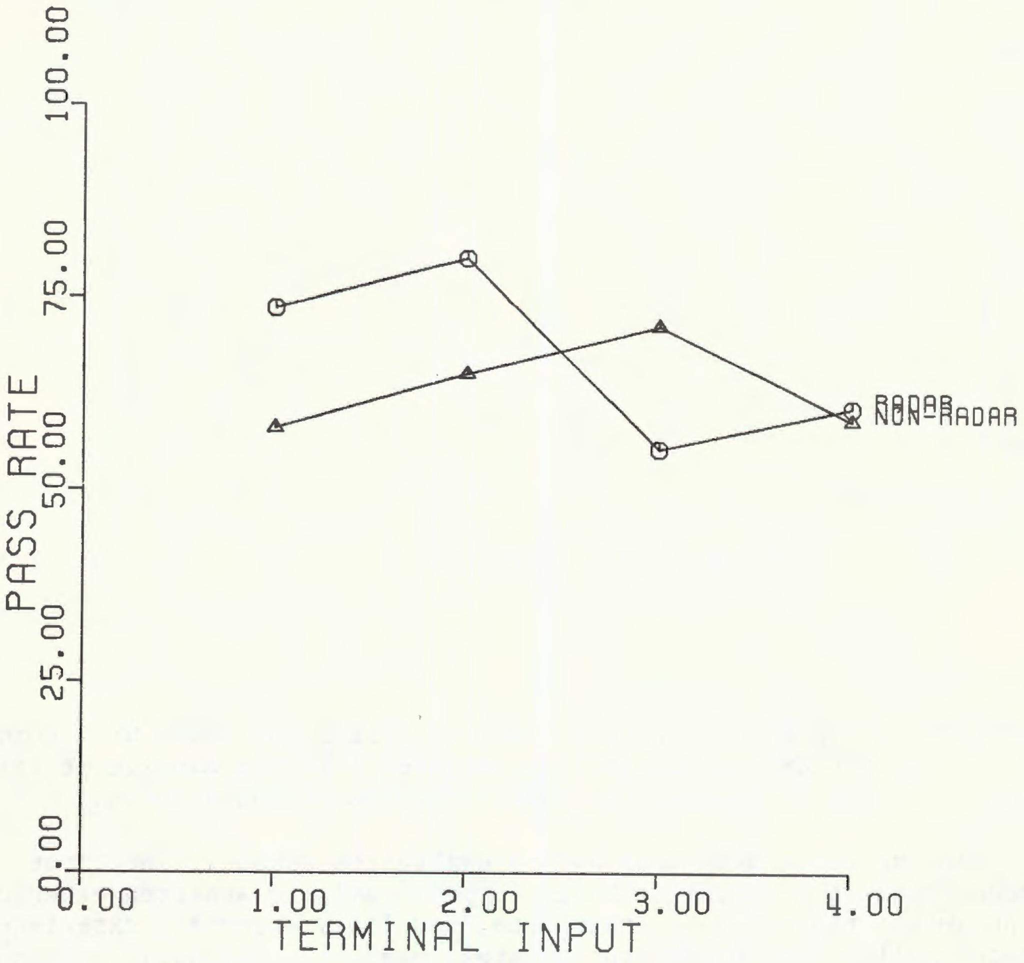

Figure 3. Interaction of Pass Rates in Nonradar and Radar Across Evaluation Phases for the Terminal Option (N=436).

This demonstrates a clear interaction effect. The en route pass rates across all intervals could be used as a textbook example for interaction. There is almost an exactly equal and opposite reaction between nonradar and radar pass rates. This effect clearly explains a substantial amount of the fluctuations in pass rates found in the two formative evaluation loops.

Parallelism of Measures. Tables 1 and 2 show the data on parallelism of laboratory problem versions. Some of the versions do not satisfy the criteria of equal means and variances to be deemed parallel. At least two factors appear to be involved in the lack of parallelism for these measures. The first is the reliability of the measures (to be discussed below). Low to moderate reliability in measurement directly affects the parallelism of measures.

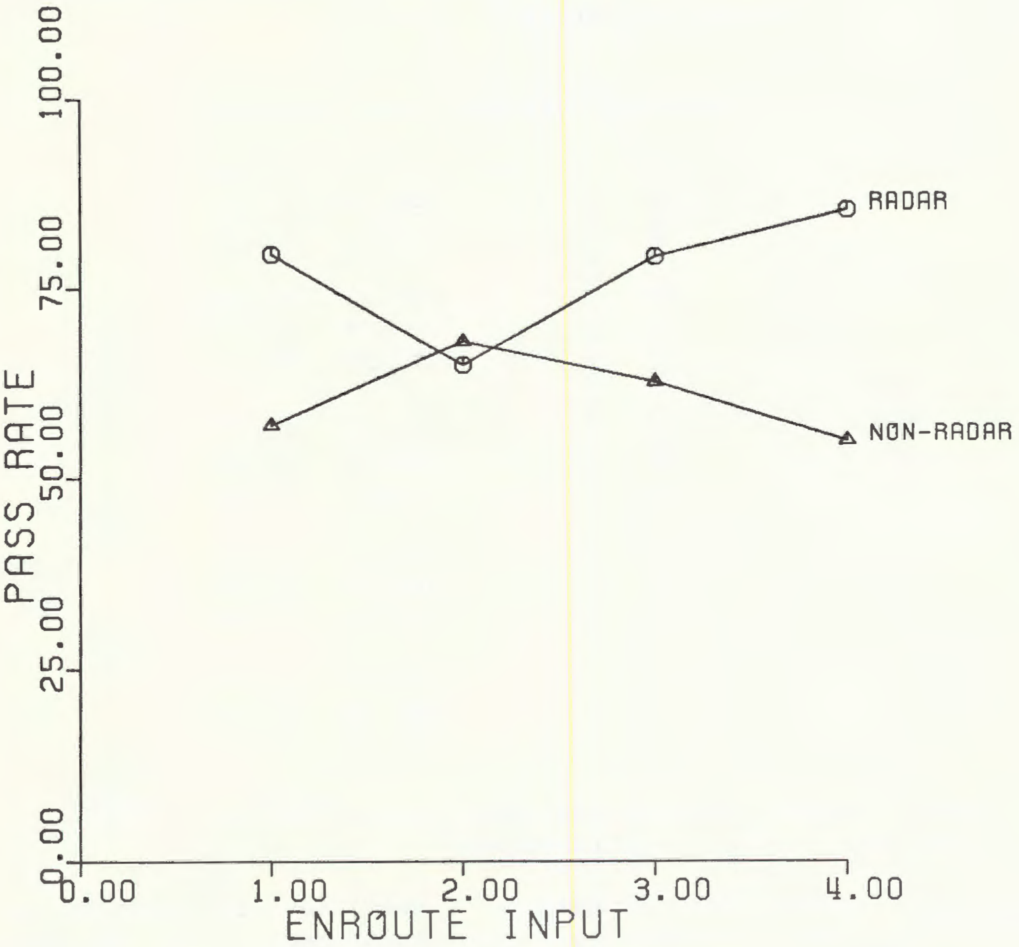

Figure 4. Interaction of Pass Rates in Nonradar and Radar Across Evaluation Phases for the En Route Option (N=328).

Since the classical definition of reliability is the correlation between two parallel measures, reliability and parallelism are interactive properties of measurement. Significant improvement in reliability should improve the parallelism of the versions. Secondly, some of the versions appear to have a ceiling effect, and this affects the equality of variances for that set of measures. As viewed in Tables 1 and 2, when the mean of a particular version is higher than the other two versions, the variance of the version with a higher mean has a smaller standard deviation. The ceiling for scores is 100, and as the means move toward 100 there is less room at the top for scores to vary. This, in turn, decreases the variance measure. Making the three versions of each problem equal in difficulty should improve the equality of the variances.

Table 1. Means and Standard Deviations for All Three Versions of Each Laboratory Problem for Terminal Formative Classes (Loop 2)

PROBLEM	VERSION	MEAN	STD DEV
1	1	57.61	17.13
1	2	56.37	17.52
1	3	58.84	20.53
2	1	69.98	15.35
2	2	69.32	19.67
2	3	71.79	18.61
3	1	73.31	17.75
3	2	*84.03	10.93
3	3	71.45	17.20
4	1	74.41	14.99
4	2	*62.80	18.29
4	3	70.65	18.44
5	1	72.75	16.05
5	2	*64.70	18.99
5	3	70.46	19.83

*Means and/or standard deviations significantly different at the $p < .05$ significance level.

Reliability. Reliability is a measure of the internal consistency of the scores obtained on the various assessment instruments. The 1975 study (4) by the Institute for Defense Analysis on ATCS selection and training reviewed several ATC studies and reported the reliability of measures used to assess ATCS performance. The review showed that coefficients from past studies ranged between .300 and .600. These are low-to-moderate coefficients. Tables 3 and 4 show, not surprisingly, that the RTF laboratory score reliabilities fall in this range.

Table 2. Means and Standard Deviations for All Three Versions of Each Laboratory Problem for En Route Formative Classes (Loop 2)

PROBLEM	VERSION	MEAN	STD DEV
1	1	77.55	10.96
1	2	76.80	14.71
1	3	80.22	6.83
2	1	*66.80	13.51
2	2	74.57	13.95
2	3	78.85	11.13
3	1	*80.67	8.98
3	2	*85.08	10.71
3	3	73.58	18.50
4	1	*63.20	14.54
4	2	71.00	12.66
4	3	77.67	11.09
5	1	66.42	10.06
5	2	71.43	14.55
5	3	70.50	11.63

*Means and/or standard deviations significantly different at the $p < .05$ significance level.

In measurement procedures, reliability is important to validity. High reliability is generally viewed as a necessary (but not sufficient) attribute for high validity. Traditionally, measurement experts have held that a measurement's validity cannot exceed its reliability. A clear discussion of this principle and the expected increase in the validity coefficients due to eliminating unreliability can be found in Lord and Novick (6). A primary purpose of the initial validation procedures (the formative evaluation loops) was to monitor and improve this necessary attribute of the program to help assure high validity. Reliability is the first building block in constructing a valid training program. Difficulty with reliability in assessing ATC performance has several sources.

One source of variation in assessing radar performance that contributes to a lack of internal consistency is the difficulty of assessing radar techniques. In nonradar where the aircraft are not visually viewed on a scope, ATC operations are more rule bound. Radar control requires more

Table 3. Intercorrelations for All Five Laboratory Problems for Terminal Formative Evaluation Classes (Loop 2)

	LS1	LS2	LS3	LS4	LS5
LS1	1.0000 (193) P=0.000	0.3438 (193) P=0.000	0.2278 (193) P=0.001	0.3096 (193) P=0.000	0.3187 (193) P=0.000
LS2	0.3438 (193) P=0.000	1.0000 (193) P=0.000	0.3959 (193) P=0.000	0.4298 (193) P=0.000	0.3250 (193) P=0.000
LS3	0.2278 (193) P=0.001	0.3959 (193) P=0.000	1.0000 (193) P=0.000	0.3304 (193) P=0.000	0.2443 (193) P=0.000
LS4	0.3096 (193) P=0.000	0.4298 (193) P=0.000	0.3304 (193) P=0.000	1.0000 (193) P=0.000	0.3694 (193) P=0.000
LS5	0.3187 (193) P=0.000	0.3250 (193) P=0.000	0.2443 (193) P=0.000	0.3694 (193) P=0.000	1.0000 (193) P=0.000

Table 4. Intercorrelations for All Five Laboratory Problems for En Route Formative Evaluation Classes (Loop 2)

	LS1	LS2	LS3	LS4	LS5
LS1	1.0000 (123) P=0.000	0.2341 (123) P=0.005	0.3288 (123) P=0.000	0.2157 (123) P=0.008	0.1601 (123) P=0.038
LS2	0.2341 (123) P=0.005	1.0000 (123) P=0.000	0.1265 (123) P=0.082	0.1426 (123) P=0.058	0.2086 (123) P=0.010
LS3	0.3288 (123) P=0.000	0.1265 (123) P=0.082	1.0000 (123) P=0.000	0.2400 (123) P=0.004	0.0736 (123) P=0.209
LS4	0.2157 (123) P=0.008	0.1426 (123) P=0.058	0.2400 (123) P=0.004	1.0000 (123) P=0.000	0.1685 (123) P=0.031
LS5	0.1601 (123) P=0.038	0.2086 (123) P=0.010	0.0736 (123) P=0.209	0.1685 (123) P=0.031	1.0000 (123) P=0.000

technique: for example, judging the turn arc of a particular aircraft traveling at a certain speed with an angling wind. CAMI's preliminary grading study (3) resulted in data relevant to this issue. Figure 8 shows that of the three major components of the laboratory technical score (procedural errors, system errors, and system deviations), procedural errors by a large margin constitute the majority of the errors. "Procedural errors" is the category in which radar control technique errors are listed. Another analysis from the study also illustrates this problem. In the process of standardizing the laboratory grading manual, profiles were calculated for each instructor by summing the frequency of errors counted by the instructor in each of the different categories for errors under procedural errors. Individual subject's laboratory problems were also profiled in the same manner. It was found that the systematic occurrence of errors in various categories was largely accounted for by instructor grading profiles. Subsequent interviews with instructors revealed that each instructor did consider some techniques more important than others and was generally more attentive to those errors: for example, the use of speed control in lieu of good vectoring procedures. The laboratory grading manual development staff utilized this information in standardizing the grading manual; however, the data do illustrate preferences for techniques. It should be carefully noted that any of the radar techniques recommended by the instructors would have led to safe and expeditious air traffic control, as evidenced by the instructors' years of successful ATC experience; however, preferences can potentially lead to differing styles of grading.

Field tracking data on post-Academy attrition due to training failure, presented in Table 5, indicate that even though the reliability of Academy measurement is only moderate, field attrition due to training failure has shown a highly significant decrease since Academy screening began, a strong indication of the effectiveness (validity) of the program. It is evident on the basis on what appears to be contradicting information, that there must be some compensating variables in the screening process that tend to offset the effects of a less than optimal reliability.

Table 5. Nonradar Academy and Field Attrition Rates, Pre- and Post-Pass/Fail

	Nonradar Academy	Field (18-24 Month Average)
Prior to Pass/Fail	---	38%
1976 Academy Pass/Fail	30%	8%

Annual Cost Avoidance Due to Academy Screening *$12.0 M

*Based on 1979 FAA Office of Management and Budget (OMB) estimates

A major interacting variable is the selection ratio. In a usual pre-strike year, there could be approximately 25,000 applicants for ATCS positions. After initial and Academic screening about 1,350 would enter long-term field training. This would result in a highly stringent 5-6 percent selection ratio. A very selective screening process is to be expected in crucial safety-related occupations requiring unique aptitudes. The effect of extreme selection on future job sucess is illustrated by Figure 5. Suppose low reliability has led to a validity coefficient of .300 which when contoured, as shown in the figure, is only slightly oval. Those persons in area I make it through to job success. Areas II and IV fail the initial screening, and areas III and IV fail on-the-job performance. Area III illustrates the comparatively small post-Academy attrition rate in the field.

Area II in Figure 5, a comparatively large area, illustrates the proportion of persons who failed initial screening but who could have been successful on the job. A comparison of areas II and III indicates that, while on-the-job training attrition in the field is small, the process used to achieve small field attrition is somewhat inefficient. A two by two matrix table, given in Figure 5, helps clarify the issue.

Analyzing the quadrants in Figure 5 reveals that another variable can be specified that affects the decision process in selection procedures, the cost of inaccurate prediction. Quadrants I and IV result in no cost due to inaccurate predictions. Quadrants II and III do result in costs for inaccuracy. A comparison of these two costs brings things into sharper focus. Applying this comparison to the Academy screening process and the question of reliability, two aspects are considered, (i) the comparative cost of failing candidates at the Academy when they would be successful in field training and passing candidates at the Academy when they would later fail field training, and (ii) how much of the inaccuracy of prediction is due to unreliable measurement.

The Academy screening program requires about 3 months, while field attrition occurs approximately 2 to 2.5 years into field training. About a 1 to 10 ratio exists between the two time frames. Considering this ratio and assuming approximately equal costs, the program could allow 10 times as many false positives as false negatives before the two costs would be equal. Further, the cost and effort to improve reliability may not merit the expenditure for the potential improvement in reliability. As stated above, reliability is a necessary but not sufficient attribute of validity; consequently, large improvement in reliability may result in only modest increases in validity. All of these factors would need to be considered.

This discussion is not intended to suggest that no effort should be made to improve reliability. It is presented as a framework both to evaluate the meaning of the reliability coefficients and to evaluate the amount of effort that should be expended to improve the efficiency of the program through better reliabililty. A complete and accurate cost analysis would be recommended prior to making judgments on whether to launch a major program to improve reliability. Such an analysis should include all of the FAA screening/training efforts, not just radar screening.

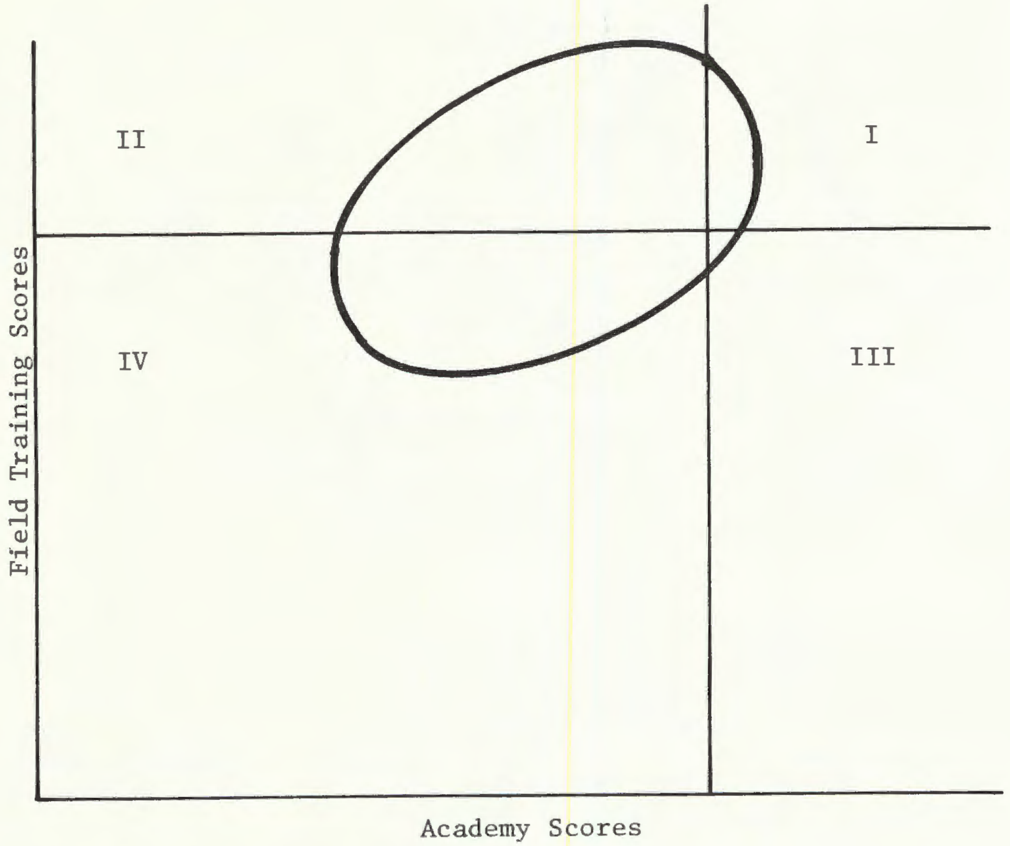

Figure 5. Effects of Extreme Selection on Future Job Attrition.

There are two recommended actions to improve reliability that require a minimal effort. (i) Perform a thorough review of the laboratory grading manual to insure that types of errors are clearly and explicitly defined with a large (within reason) number of examples offered. The general problem in reliability is internal consistency. If two separate styles of handling aircraft result in safe and expeditious aircraft movement in accordance with the FAA 7110.65 manual, then it should not be the case that one instructor counts the procedure as an error (presuming the instructor views it as a less than optimal technique) while another instructor views the procedure as correct. Differing interpretations of this sort are what lead to unreliability. (ii) Another recommendation that may improve reliability involves instructor training. This recommendation was first made as a result of the 1979 computer-based instruction (CBI) study on the pilot/ghost study. A set of scenarios with good illustrations of the errors listed in the laboratory grading manual should be systematically employed to train instructors in assessing a candidate's performance. Instructors should not be released from training until they can demonstrate the ability consistently to grade correctly on these canned scenarios. The process should be a well designed CBI component of instructor training.

The second set of suggestions on reliability is for future study and consideration. CAMI's preliminary grading study (3) indicated that some of the automated measures showed promise for assessing student performance. Development work on data reduction software is presently underway at the Aeronautical Center Data Services Division. (i) It is recommended that a full study be made of the automated measures to assess their use in RTF grading. Computer measures of performance are almost absolutely reliable on a test/retest basis. In 1978, CAMI and the validation committee, based on site visits where very little standardized, objective measurement was found, recommended the use of a grading unit in the RTF that would be responsible for the technical assessment portion of laboratory grading. The instructor observing the running of the problem would still be responsible for the instructor assessment portion of laboratory grading. In this manner a blind review of problems on replay would occur with systematic reliability checks built into the process. A cursory review of this suggestion in 1978 revealed several manpower problems current at that time. (ii) It is suggested that a thorough review be made in relation to possible gains from the approach.

The Issue of Fairness. The Uniform Guidelines on Employee Selection (7) state that any assessment process used as a basis for decisions affecting employee status must be shown to be fair. The ultimate test of fairness is based on predictive studies that employ an external on-the-job success measure. Briefly stated, the probability of selection must be equal to the probability of success on the job without regard to minority group or sex. A concise review of fairness models can be found in Lewis (5). At present, an external criterion, i.e., on-the-job success information, is not available on RTF trainees. As a result, satisfaction of the Uniform Guidelines test for fairness cannot be reviewed until the field tracking data have been collected and summative evaluation occurs. However, the guidelines do state a criterion for a beginning program to use in determining if an external criterion-related study is required. It is termed the "four/fifths rule." This rule asserts that the selection rate for any minority group should not be less than four-fifths of the selection rate for the majority group. Figures 6 and 7 show the various pass rates for nonminority men (the majority group) and for racial minorities and women across the evaluation phases. These numbers indicate that a criterion-related study of fairness is required. This study will be done as a part of the summative, field tracking, evaluation phase.

In the interim there are some recommended actions to assist in the fairness of the program. These actions are in the form of reviews. It is recommended (i) that sufficient data be collected to thoroughly study fairness in the summative evaluation, (ii) that fairness be included as an intrinsic part of the study on automated measures (the computer calculates the same without regard to minority status or sex), and (iii) that the problem of fairness be assessed in terms of a special RTF grading unit, since a blind review of a scenario would not reveal the minority status or sex of the person being assessed.

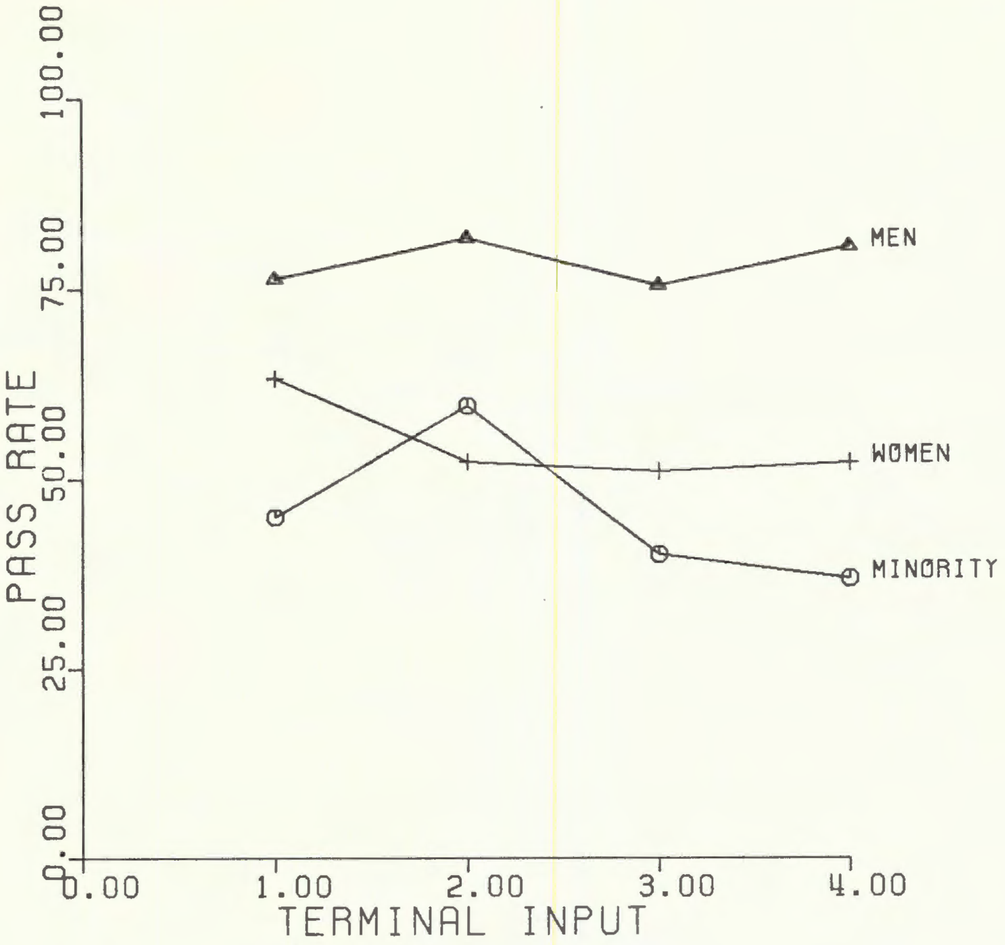

Figure 6. Pass Rates Across Evaluation Phases for Nonminority Men, Women, and Minorities for the Terminal Option.

Computer Outages During Graded Problems. A complete RTF laboratory problem is 30 minutes in duration. Some RTF inputs have experienced computer outages in as high as 10-12 percent of their graded problems prior to the completion of the problem. The overall average is nearer to 5-7 percent. Recent improvements in the computer hardware have resulted in a significant improvement in computer performance; however, a complete assessment of how often (on the average) outages will occur during graded problems, given the improvements, has not been done. The primary issue does not center on whether outages can be eliminated. After all, computers are imperfect and consequently subject to failure. The issue involves what action is to be

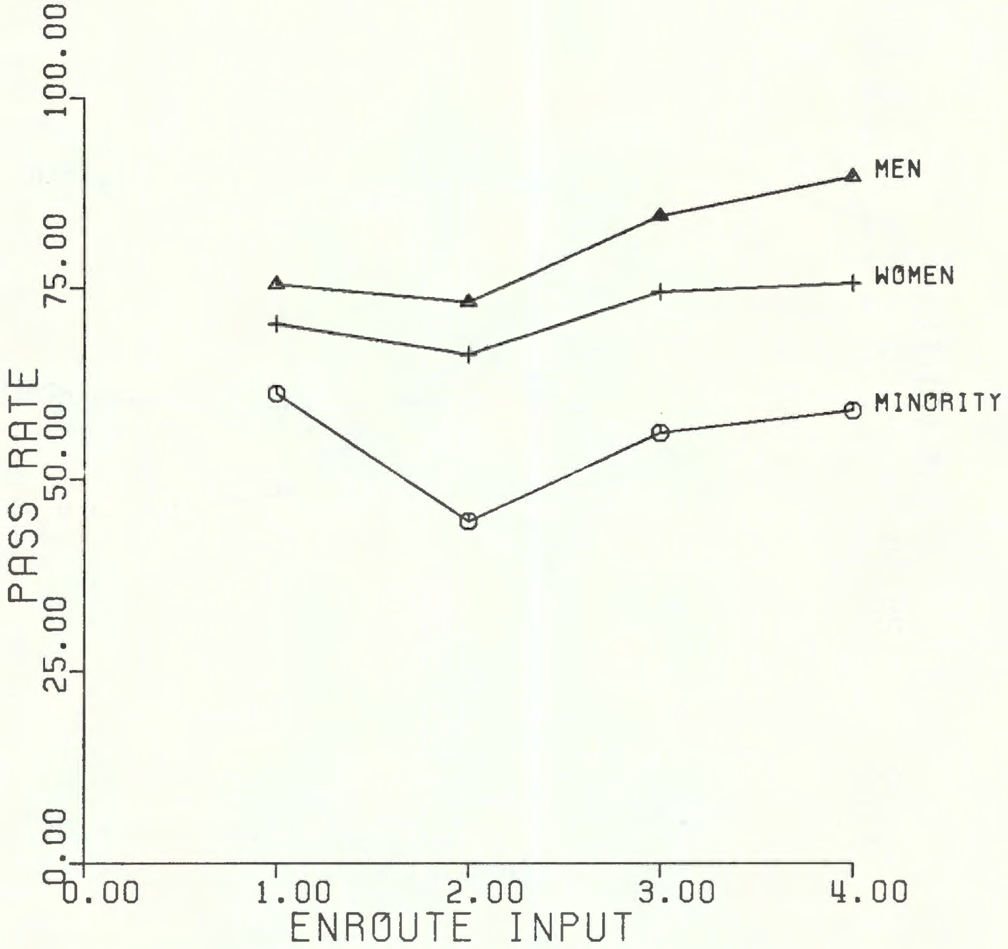

Figure 7. Pass Rates Across Evaluation Phases for Nonminority Men, Women, and Minorities for the En Route Option.

taken with regard to failures during graded problems. Figure 8 is a graph showing the frequency of errors distributed across minutes in the problem. The graph demonstrates an increase in the density of errors as the problem time increases. This is expected since the problem's complexity grows as the problem proceeds. The basic question involved is, "How long should the problem run before it is deemed a full problem?" Is 20 minutes sufficient? 25 minutes? A decision on the matter must also include a consideration of the consequences. Deeming that a graded problem run is insufficient in duration means that another version of the same problem must be administered. Does this give the candidate taking a rerun an unfair advantage? What if the candidate performs far worse on the rerun than he performed on the original run?

Table 6 contains a decision matrix with the consequences to each group, based on the effects the decision will have on total laboratory score, if a rerun is given for a less than completed problem and if problems of a stated duration less than the full 30 minutes are accepted.

Table 6. The Effects of Accepting Less Than a Full 30-Minute Run and Allowing a Problem Rerun for Candidates Involved and Not Involved in Incomplete Problem Runs.

	Effects of accepting less than full run (25 minutes or more)	Effects of allowing a rerun	
		Does better on rerun	Does worse on rerun
Candidates involved in an incomplete problem run	Advantage	Advantage	Disadvantage
Candidates not involved in an incomplete problem run	No Effect	No Effect	No Effect

For candidates whose problems run 25 to 29 minutes, accepting those runs will have an advantageous effect, since there is less opportunity for errors to occur. If a rerun is performed, a better performance on the rerun is advantageous; a worse performance is not advantageous. The global conclusion for this group would be an advantageous effect. For the group not involved in incomplete problems, since pass/fail is criterion referenced (having a cut-point score) as opposed to norm referenced (ranking persons by scores and selecting a percentage of the top scorers), there is no effect on their total laboratory score or their probability of passing or failing. Consequently, allowing reruns for less than complete problems and specifying a minimum number of minutes required for an acceptable problem would be a reasonable approach. Viewing the error distributions across time, Figure 8, shows that about 95 percent of the errors occurred by 25 minutes into the problem. The sharpest increase in system errors and procedural errors also occurred at about 25 minutes and subsequently leveled off. Based on this information, the recommer lation in regard to computer outages is as follows: A 25-minute minimum be required to qualify as a valid problem. If the problem runs less than 25 minutes prior to an outage, the candidate should be rescheduled to run another version of that same problem until a valid run is obtained.

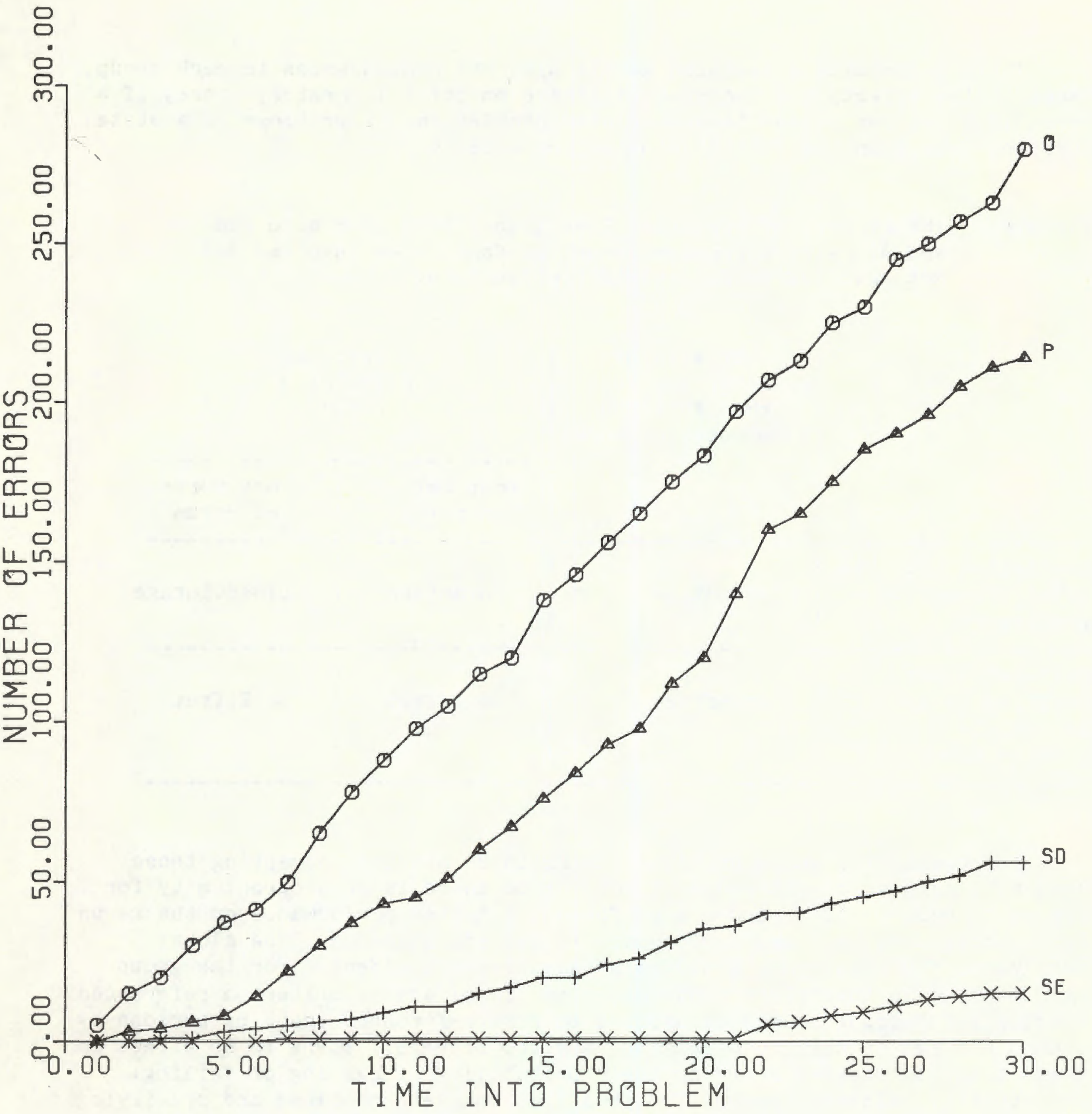

Figure 8. Distributions of Types of Laboratory Errors Across Minutes in the Problem.

Concurrent Validity Correlations. Three primary measures are involved in the RTF phase score calculation used for pass/fail, (i) CPT, (ii) CST, and (iii) laboratory scores. The same is true for the nonradar training phase. This offers a means of reviewing the concurrent validity of the two phases with a multitrait/multimethod approach. Tables 7 and 8 contain the correlation matrix of radar and nonradar CPT, CST, and laboratory scores. The multitrait/multimethod approach has the expectation that internal measures should correlate highly, while measures not internal but similar in trait should correlate moderately. The overall correlation between the radar and nonradar phase composite score is .402, indicating that a similar trait is being measured but the scores are not duplicative.

The correlations for the nonradar data (NCPT, NCST, NLA) are slightly restricted (lower) since only passing nonradar candidates (those with composite scores above 70) are in the sample. However, the correlations for both terminal and en route demonstrate that the radar and nonradar programs have a proper concurrent relationship. In general, correlations of internal program measures are higher than with external measures, with the external correlations being high enough to assume that a similar trait is being measured. The only notable exception is in terminal where the internal nonradar correlation for laboratory average and CST is -0.0216. A negative correlation in this restricted group, however, is not surprising since this correlation has typically been quite low in the CAMI long-range data base. Several correlations demonstrate the concurrent validity of the radar

Table 7. Correlation Matrix of Radar and Nonradar CPT, CST, and Laboratory Scores for Terminal.

	MCPT	MCST	AVL5	NCPT	NCST	NLA
MCPT	1.0000 (440) P=0.000	0.3024 (440) P=0.000	0.2392 (435) P=0.000	0.2757 (440) P=0.000	0.1881 (440) P=0.000	0.1243 (383) P=0.007
MCST	0.3024 (440) P=0.000	1.0000 (440) P=0.000	0.1876 (435) P=0.000	0.1318 (440) P=0.003	0.2515 (440) P=0.000	0.2209 (383) P=0.000
AVL5	0.2392 (435) P=0.000	0.1876 (435) P=0.000	1.0000 (435) P=0.000	0.1801 (435) P=0.000	0.1703 (435) P=0.000	0.3434 (379) P=0.000
NCPT	0.2757 (440) P=0.000	0.1318 (440) P=0.003	0.1801 (435) P=0.000	1.0000 (440) P=0.000	0.2969 (440) P=0.000	0.2194 (383) P=0.000
NCST	0.1881 (440) P=0.000	0.2515 (440) P=0.000	0.1703 (435) P=0.000	0.2969 (440) P=0.000	1.0000 (440) P=0.000	-0.0216 (383) P=0.337
NLA	0.1243 (383) P=0.007	0.2209 (383) P=0.000	0.3434 (379) P=0.000	0.2194 (383) P=0.000	-0.0216 (383) P=0.337	1.0000 (383) P=0.000

Table 8. Correlation Matrix of Radar and Nonradar CPT, CST, and Laboratory Scores for En Route.

	MCPT	MCST	AVL5	NCPT	NCST	NLA
MCPT	1.0000 (298) P=0.000	0.6494 (298) P=0.000	0.2664 (296) P=0.000	0.2804 (298) P=0.000	0.2114 (295) P=0.000	0.1965 (262) P=0.001
MCST	0.6494 (298) P=0.000	1.0000 (298) P=0.000	0.3720 (296) P=0.000	0.3519 (298) P=0.000	0.2180 (295) P=0.000	0.1367 (262) P=0.013
AVL5	0.2664 (296) P=0.000	0.3720 (296) P=0.000	1.0000 (296) P=0.000	0.2694 (296) P=0.000	0.2918 (293) P=0.000	0.2161 (261) P=0.000
NCPT	0.2804 (298) P=0.000	0.3519 (298) P=0.000	0.2694 (296) P=0.000	1.0000 (298) P=0.000	0.2553 (295) P=0.000	0.1572 (262) P=0.005
NCST	0.2114 (295) P=0.000	0.2180 (295) P=0.000	0.2918 (293) P=0.000	0.2553 (295) P=0.000	1.0000 (295) P=0.000	0.2041 (262) P=0.000
NLA	0.1965 (262) P=0.001	0.1367 (262) P=0.013	0.2161 (261) P=0.000	0.1572 (262) P=0.005	0.2041 (262) P=0.000	1.0000 (262) P=0.000

program. For example, the correlations between radar and nonradar CPT are higher than the correlations of radar CPT with nonradar CST, while the radar CPT and radar CST correlation is higher than radar CPT and nonradar CST. The same pattern holds true through both matrices. It does appear in the en route matrix that the CPT and CST are still too duplicative, as was found earlier. A more thorough review of the two tests should be undertaken. Based on the multimethod/multitrait correlation matrix, the initial RTF program, overall, shows sufficiently high concurrent validity.

Data Reduction and Analysis (DR and A). At present, DR and A is in software development, so a full assessment is not feasible. The preliminary CAMI grading study (3) indicated some promise for its use; however, machine-scoring in a complex interactive system such as ATC must be approached with considerable caution. At this point it is believed that a machine-scoring system will require a large degree of human interfacing to produce optimal results. The positive features of a potential machine-scoring system in relation to measurement reliability and fairness were previously discussed. At this point it appears sound advice to continue a program of assessing DR and A as a potential enhancement in RTF assessment, but not with the unrealistic expectations that human scoring can be totally eliminated. DR and A scoring should become more feasible as the ATC system moves toward more extensive automation in the 1990-2000 period.

Initial Evaluation Conclusion. Few training programs undergo such minute detailed scrutiny as has the ATCS Academy programs. According to OMB estimates, the present Academy screening program, based on nonradar, results in an annual cost avoidance of about $12 million. The success of this program is in large part due to a thorough evaluation process and the commitment of the FAA to act on the basis of data-based decisions in structuring the Academy programs. In programs involving public safety nothing less will suffice. The RTF initial evaluation has been detailed and thorough. While the data collected in this initial phase indicate some areas for concern (the parallelism and reliability of laboratory scores and the fairness of the measures in general), these concerns must be realistically evaluated with the perspective that the data are based on measures collected while the program was not pass/fail, and, as any ATCS or pilot will vouch, conditions change in moving from simulation to the real thing. The data in general show that the program is headed toward the target. To overinterpret simulation data as very precise would be a mistake. It is the judgment of the CAMI evaluation staff that it is time to begin the pass/fail program and make future refinements on the basis of actual data.

REFERENCES

1. Boone, James O. 1982. A generic model for evaluation of the Federal Aviation Administration air traffic control specialist training programs. FAA Office of Aviation Medicine Report No. FAA-AM-82-2, Washington, D.C.

2. Boone, James O. 1980. System performance, error rates, and training time for recent FAA Academy nonradar graduates, community persons, and handicapped persons on the radar training facility pilot position. FAA Office of Aviation Medicine Report No. FAA-AM-80-5, Washington, D.C.

3. Boone, James O. 1980. The Federal Aviation Administration's Radar Training Facility and employee selection and training. FAA Office of Aviation Medicine Report No. FAA-AM-80-15, Washington, D.C.

4. Henry, J.H. 1975. Training of U.S. air traffic controllers. Institute for Defense Analysis, Report R-206, Arlington, Virginia.

5. Lewis, Mary A. 1979. A comparison of three models for determining test fairness. FAA Office of Aviation Medicine Report No. FAA-AM-79-3, Washington, D.C.

6. Lord, F.M. and M.R. Novick. 1968. Statistical theories of mental test scores. Addison-Wesley Publishing Co., Reading, Massachusetts.

7. ______. 1978. Uniform guidelines on employee selection procedures. Federal Register, Vol. 43, No. 251, p. 38290, August.