TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FAA-AM-72-5		

4. Title and Subtitle	5. Report Date
MULTIPLE TASK PERFORMANCE AS A PREDICTOR OF THE POTENTIAL OF AIR TRAFFIC CONTROLLER TRAINEES	January 1972
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
W. Dean Chiles, Ph.D., Alan E. Jennings, B.A. Georgetta West, B.S. (with the assistance of William T. Abernathy)	

9. Performing Organization Name and Address	10. Work Unit No.
FAA Civil Aeromedical Institute P. O. Box 25082 Oklahoma City, Oklahoma 73125	11. Contract or Grant No.
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address Office of Aviation Medicine Federal Aviation Administration 800 Independence Avenue, S. W. Washington, D. C. 20591	OAM Report
	14. Sponsoring Agency Code

15. Supplementary Notes

This research was performed under Task AM-A-71-PSY-29.

16. Abstract

Two hundred and twenty-nine Air Traffic Controller trainees were tested on the CAMI Multiple Task Performance Battery. The battery provides objective measures of monitoring, arithmetical skills, visual discrimination, and group problem solving. The criterion of trainee potential was based on ratings from FAA Academy instructors in courses being attended by the trainees. Five studies were conducted, the first one (N = 19) being in the nature of a pilot study for checking out procedures. The second study (N = 60) yielded a validity coefficient of .54. The third study (N = 31) yielded a coefficient of .53. The fourth study (N = 30) found no predictive power for the MTPB. The fifth study (N = 89) produced a coefficient of .24 for one method of computing the performance index and .46 for a second method. For each study, the coefficient is based on one hour of testing with about 50 minutes of preceding instruction and practice. It is concluded that the MTPB-approach to selection offers promise as a screening device for Air Traffic Control Specialist applicants, but further research is required to establish this as a fact and to determine its utility in terms of cost-effectiveness.

17. Key Words	18. Distribution Statement
Selection Complex Performance	Availability is unlimited. Document may be released to the National Technical Information Service, Springfield, Virginia 22151, for sale to the public.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	31	$3.00

Form DOT F 1700.7 (8-69)

MULTIPLE TASK PERFORMANCE AS A PREDICTOR OF THE POTENTIAL OF AIR TRAFFIC CONTROLLER TRAINEES

I. Introduction.

The Civil Service Commission Air Traffic Control Specialist Screening Battery does an efficient job of identifying those individuals who possess the "elemental" skills and knowledges necessary to becoming a satisfactory controller (Trites and Cobb, 1964). However, it does not measure a particular kind of ability that may be said to be a defining characteristic of a good controller—the ability to perform several different tasks "simultaneously." Thus, undoubtedly a number of trainees are eliminated from Air Traffic Service training programs, not because these trainees lack information in the academic sense, but because of an apparent deficiency in their skill at the concurrent, time-shared performance of a variety of tasks. They cannot "put this information together" and perform satisfactorily. If this is true (and it is a widely held belief among FAA Academy instructors), then a technique that provides an objective, reliable index of the ability of an applicant to time-share multiple tasks of relevance to the controller's job would be of value as an adjunct to the existing selection devices.

The Multiple Task Performance Battery (MTPB) was originally designed for the Air Force as a device for use in research on complex performance of the sort demanded of air crew personnel (Chiles, Alluisi, and Adams, 1968). The elements of the MTPB were selected to provide objective measures of "psychological or behavioral functions" of relevance to Air Force operations. The functions measured by the MTPB include monitoring, information processing, mental arithmetic, visual discrimination, and inter-individual interaction in the execution of procedures. These functions would appear to be relevant not only to air crew activities but to complex jobs in general and to the job of the Air Traffic Control Specialist in particular. Moreover, the tasks, as routinely used over a number of years, have been structured to impose varying levels of demand on the individual with respect to the requirements for time-sharing. Good performers on the MTPB are those individuals who not only possess ability on the individual tasks but, also, are readily able to shift their focus of attention from one kind of activity to another without disruption of the on-going process.

The rationale for investigating the selection potential of the MTPB derives from the complexity of the task situation with respect to the requirements for time sharing.* There is no particular reason to expect any of the individual tasks to be predictive of Air Traffic Control Specialist abilities even though, as will be suggested in the results section of this report, FAA Academy instructors regarded the tasks as being generally of relevance to the job of the controller. The nature of the MTPB is such that an individual who fixates on the performance of one task will almost inevitably do poorly on one or more of the time-shared tasks. This is especially evident early in learning when skill is being acquired at the detailed, refined performance of the individual tasks. Thus, during this first hour or so of learning, the task ensemble should be sensitive to differences in the ability of individuals to shift attention rapidly from one task

* Very general planning for such an undertaking was initiated about a year before the studies to be described here got underway. Appropriate liaison with the FAA Academy materialized in the form of an Academy instructor who served as a volunteer in another experiment being conducted by CAMI. This instructor was struck by what he thought to be the similarities between the demands placed on him by a multiple task device (that differed somewhat from the MTPB) and the demands placed on him as a journeyman En Route controller. This chance event paved the way for the study and, in a number of significant respects, made the ultimate scope of the study possible.

to a second task and to quickly adopt the proper "set" for the performance of that second task.

The intent of the studies was to provide an answer to an intermediate question. We were not in any sense attempting to develop the quantities of supporting data that would be necessary to justify the formal implementation of the MTPB as a selection device for screening Air Traffic Control Specialist applicants. The intermediate question we were trying to answer concerned the determination of whether or not the *approach* embodied in the MTPB showed sufficient promise as a selection device to warrant a large-scale investigation to develop the supporting data that would be required to demonstrate its efficiency as a predictor of controller skills in terms of cost-effectiveness criteria.

The methodology that we had previously used with the MTPB was, broadly speaking, applicable, but it had to be modified to take account of a new set of constraints imposed by the selection paradigm. For the most part, the nature of our previous research with the MTPB (Chiles, Alluisi, and Adams, 1968) had not involved any very stringent restrictions with respect to the time required for mastery of the tasks. With the task combinations and task difficulties used in those studies, we had established that the average individual required between 16 and 20 hours of practice before his performance leveled off. And with one of the tasks, mental arithmetic, many individuals demonstrated very little progress in mastering the tasks until after about 30 minutes of practice. Thus, it was clear that it would not be reasonable to use the specific task structures that we had previously employed. On the other hand, because we had amassed large amounts of data on the tasks with a fixed set of difficulties, we wanted to stay as close to the previously used conditions as feasible in order to make those masses of background data maximally applicable. Three temporal factors entered into the selection of the specifics of the methodology to be used in the present studies. First, the task structures must be such that, in a short time, the individual could achieve an adequate level of understanding as to how the tasks worked. Second, the testing time must be long enough to assure that the numbers of measures would be sufficient to yield adequate reliabilities of the resultant data. And third, as with most complex situations, the apparent task characteristics might very well change with increasing skill in the performance of the task complex. These times had to be weighed against the costs of administration of the test in relation to the potential cost-effectiveness of the procedure for selection purposes. Thus, because of these constraints (that had not been a problem in our previous studies), an important part of the present work was the evaluation of significant extensions of the existing methodology of complex performance measurement.

The ultimate goal of any selection enterprise is a pragmatic one—the prediction of on-the-job performance. This, of course, implies the necessity of discovering or developing criteria that are in fact descriptive of the quality of on-the-job performance exhibited by qualified specialists. If such criteria were to exist for the job of the Air Traffic Control Specialist, any study that examined the characteristics of trainees would have to wait the approximately two years it takes for an entering trainee to become a fully-qualified controller. Typically, studies of selection have dealt with intermediate criteria (Wallace, 1965). They turn out to be based on performance in some phase of the training program. Independent of the prediction of on-the-job performance, a training criterion has obvious direct practical value in that the individual must be able to get through the training program in order to become qualified, and, if you can predict those who will and will not succeed in training, substantial cost savings can be expected. An adjunct to this approach was used by Brokaw (1959) in an initial investigation of the use of paper-and-pencil selection tests for screening of Air Traffic Control Specialists. In addition to using grades achieved in the training program at the FAA Academy, Brokaw asked instructors to fill out a rating form that was designed to elicit their opinion of the potential of trainee to develop into fully–rated controllers. One to two years later, he used essentially the same form to get evaluations of supervisors as to the on-the-job performance of these trainees. One of the significant results of those studies that is of direct relevance to our criterion problem is the fact that the evaluations of the instructors at the end of the training course provided the best single predictor of the subsequent evaluations of supervisors ($r = +.59$). Trites (1961) did a follow-up study on these same trainees approx

mately five years after they had completed the training course. Trites concluded that "Instructors in the air traffic control school can make exceptionally valid predictions of (supervisor) job performance evaluation some years later." The correlation between instructor ratings and supervisor evaluations five years later was +.45. Supportive of Trites' conclusion is the fact that the 1956 instructor ratings were better than the 1957 supervisor evaluations as a predictor of the 1961 supervisor evaluations of on-the-job performance ($r = +.45$ vs. $r = +.33$). The timely reporting of this research precludes the present use of an on-the-job criterion, but both Brokaw's and Trites' findings suggest that, in the absence of a "better" criterion, instructor ratings give a good indication of the potential of the trainee to develop into a fully-rated controller. (Future follow-up of the trainees tested in this effort is planned.)

Another approach to the evaluation of a selection technique is to examine the content validity of the device. Application of this approach simply involves having a number of "experts" in the field become familiar with the device. The experts then indicate the extent to which they think the device appears to provide measures of relevance to the job or jobs in question. Although this is a time-honored procedure, it generally cannot be quantified to provide estimates of the reliability and validity of the expert opinions involved. One has to base his judgment of the validity of a specific finding on the amount of confidence he has in the ability of the experts to analyze the device in relation to the job requirements. We attempted to evaluate the content validity of the MTPB with respect to the job of the Air Traffic Control Specialist by training and testing instructors from the FAA Academy. Upon completion of the testing, the instructors were asked to fill out an evaluation form on which they indicated their reactions to various points concerning the relevance of the MTPB to the job of the Air Traffic Control Specialist. We hoped that this would provide evidence concerning the content validity of the MTPB.

II. Method.

Subjects

Study I. The 19 subjects in this study were entry-level En Route trainees who were approximately midway in the En Route controller training program at the FAA Academy. These subjects were recruited to participate in a "study on complex performance; how many dials and lights can a person monitor at one time?" They were told nothing about our interest in selection of controllers. These subjects participated on a volunteer basis during their non-duty hours. Their motivation in participating was unknown but was presumably comprised of some combination of curiosity and a desire to contribute to an FAA research effort. As with all of the groups, they were assured that the data on individuals would be available only to CAMI research personnel.

Study II. A total of 65 subjects participated in Study II; usable data were collected on 60 of them. These subjects had all completed the 9–week, En Route training course at the Academy and had returned for an interim radar course. They participated on a semi-volunteer basis. Namely, we determined that grade sheets from the Academy course were available and then asked trainees, by name, to volunteer. Only a very few trainees declined to participate. The subjects in this and the subsequent studies were released from class for the testing periods. They were given a brief but specific orientation as to the intent of the program and were told that the results with the first group were promising as regards the use of the performance measures to predict Academy training performance.

Study III. Study III involved the testing of 35 subjects who were selected from the En Route training program in essentially the same manner as were the subjects in Study II. Usable data were obtained on 31 of these men. They, too, were released from classes to participate, and they were also given the briefing regarding the intent of the study.

Study IV. Study IV involved 30 trainees from the Academy Terminal training program. They were tested during their normal duty hours within the first week of their arrival at the Academy for their 9-week training course. They were given the same briefing about the intent of the study and the disposition of the data.

Study V. Study V involved an entire class of 89 entering trainees. They were tested during the first week of the 9–week Terminal training

course. These subjects were not given a stated option with respect to participation; thus, they were not volunteers. They were given the same briefing as the previous subjects.

Instructor Study. In addition to the above studies, a total of 40 instructors from the FAA Academy were tested. Ten of these subjects were instructors in the En Route training program who were tested for a total of four hours. The remaining 30 subjects were instructors in the Terminal training program who were tested for only one hour. They were given the same general briefing as the trainees with respect to the intent of the main studies and were given the same training. They were told that we would be asking them to complete an evaluation questionnaire (Appendix I) at the end of the test session.

Apparatus and Procedure

The apparatus consisted of the CAMI Multiple Task Performance Battery (Figure 1) which has been described in detail elsewhere (Chiles, Alluisi, and Adams, 1968). The battery is comprised of five tasks, namely, monitoring a static process (warning lights), monitoring a dynamic process (probability monitoring), mental arithmetic, pattern discrimination, and group problem solving (code-lock). A brief description of the functional requirements of the tasks follows.

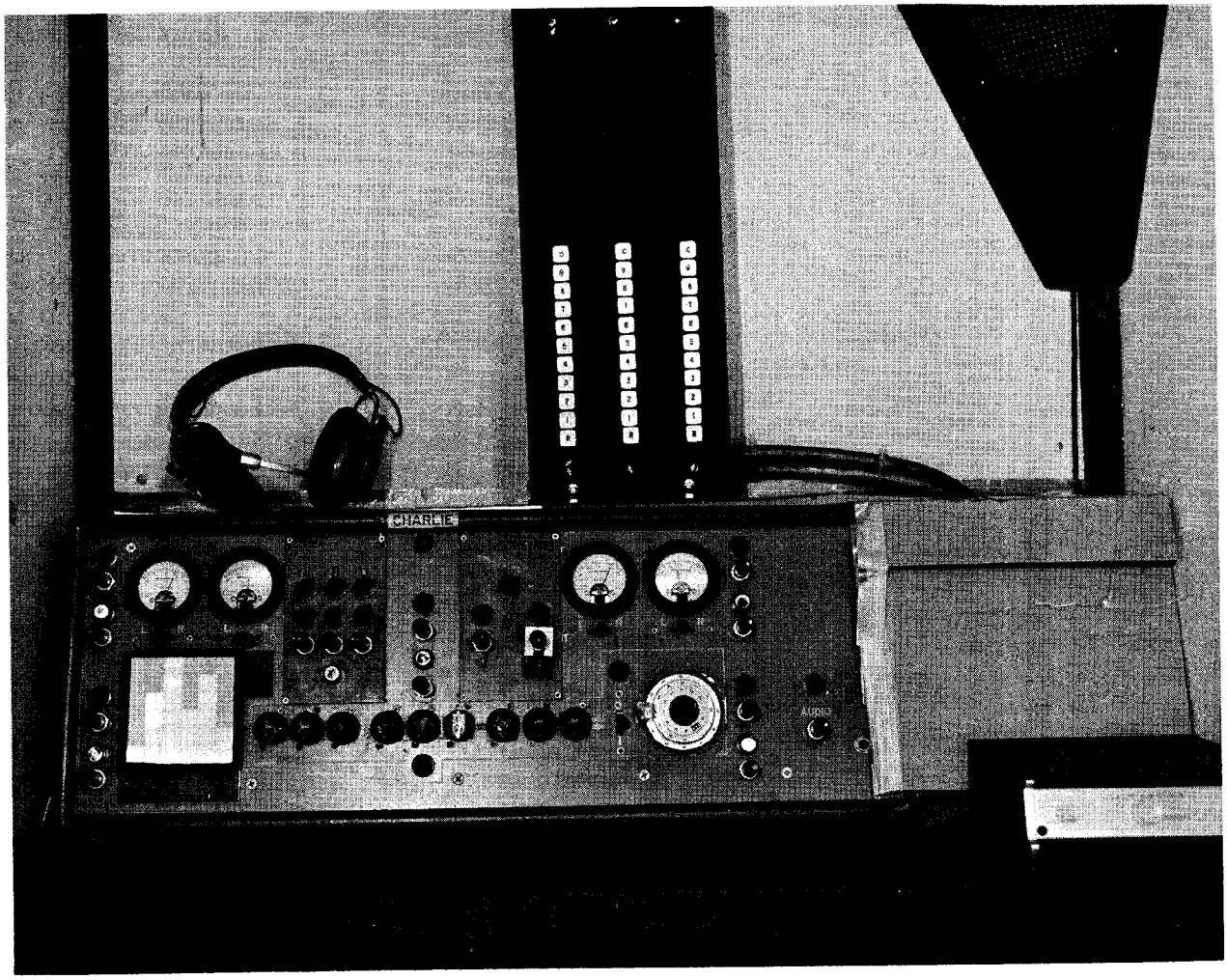

FIGURE 1. The CAMI Multiple Task Performance Battery.

Warning Lights. This task requires the subject to return any one of five green lights and five red lights to its normal state by pushing a button associated with each individual light. The normal state was for green lights to be on and red lights to be off. In Studies I through IV, an average of 72 signals (red plus green) was presented in each one-hour test session. In Study V, an average of 106 signals was presented. Separate response-time measures were recorded for the red lights and for the green lights.

Probability Monitoring. This task requires the subject to return the pointer of any one of four meters to zero (12 o'clock) when that pointer shifted to either a plus 25 or a minus 25, i.e., either one o'clock or eleven o'clock. The response was made with a three position lever switch located one below each meter. In Studies I and II, the pointer was stationary except when a signal caused it to move to and maintain a new position. In Studies III, IV, and V, the pointer wandered at random about zero with a maximum excursion of plus or minus 25; in these later studies, a signal shifted the mean position of the pointer to plus or minus 25 and the subject had to deduce that an extreme fluctuation was produced by a signal. In Studies I, II, and V, an average of 110 signals was introduced in each hour of testing. In Studies III and IV, an average of 80 signals was introduced in each hour of testing. Response time was recorded on this task, and, in addition, both false responses and missed signals were recorded, but only in Study V was the missed-signal measure used in the analysis.

Arithmetic. This task required the subject to compute the sum of two 2-digit numbers and subtract a third 2-digit number from that sum. Answers were entered in a push-button response device similar to a full-keyboard adding machine. In Studies I, II, and V, the problem elements were 2-digit numbers falling between 10 and 21, e.g., 20+19−13=26. In Studies III and IV, the problem elements were numbers between 10 and 99, e.g., 55+72−91=36. The main difference in the problems is that for the smaller element problems, subjects can generally get the answer directly; for the larger element problems, subjects have to get the answer one digit at a time. Problems were presented at the rate of three problems per minute. The actual time for which a given problem was present was about 17 seconds. Both solution time and accuracy were recorded on this task.

Pattern Discrimination. The problems on this task were analogous to a question on a multiple-choice examination. The subject was presented a "standard" pattern which looked like a vertically-oriented bar graph with each column height from one through six represented just once. This standard pattern was followed by two comparison patterns; the subject was to give one of three answers; namely, that the first comparison pattern was the same as the standard; the second comparison pattern was the same as the standard; or, neither comparison pattern was the same as the standard. The standard pattern appeared for five seconds and each comparison pattern was presented for two seconds. Problems were presented at the rate of two per minute. The performance measure used on this task was number correct.

Problem Solving. Each subject's test panel was equipped with a push-button switch and three "feedback" lights. The task required the five subjects tested at a given time to discover the correct sequence in which to press the buttons in order to turn on a green light which signified that the problem had been solved. An amber light showed that one of the buttons had been pushed and that the response had registered in the scoring system. A red light provided error information. The subjects were instructed to follow a standard search procedure, always beginning with the subject designated as Alpha and continuing through the subject designated as Echo; this procedure was used in that it involved a readily learnable search sequence that would reduce the memory load on the individual subjects and facilitate the avoidance of redundant, erroneous responses. The subjects communicated with each other via a "hot-wired" earphone system. The initial illumination of the red light indicated to the subjects that a problem was being presented and that they should immediately begin to search for the solution. Subsequently, the red light provided error information in the following manner: Any time any one of the subjects pushed his button, the red light would go out. The subjects were instructed that each man should call out his position as he pushed his button. If the button depressed was the correct first button in the problem sequence present

at that time, the light would remain out when the button was released; if the button was incorrect, the red light would come back on when the button was released. Thus, the initial step in solving a problem was to try the buttons one at a time in an Alpha through Echo sequence to determine which subject was first for solving the problem at hand. Having found that button, the search was continued in an Alpha through Echo sequence to find the next correct button. Thus, if the next man to push his button (after finding the first correct button) was the second in the sequence, the red light would stay out; if he were not next, the red light would come back on when he released his button and the number-one man for that problem would have to push again in order to resume the search for the number-two man in the sequence. The search for the remaining sequence proceeded in the same manner.

Once the correct sequence had been entered, the green "problem-solved" light would come on. For Studies I through IV, the light would remain on for 30 seconds; for Study V, it would remain on for 15 seconds. At the end of the 30-second (or 15-second) interval, the red light would be re-illuminated indicating that the previously-solved problem was again present. The subjects were to enter that same solution as rapidly as possible. The green light would again come on for the appropriate interval, and the next time the red light came on it would indicate that a new problem was present. The actual number of problems presented on this task depended upon how quickly the subjects solved individual problems since each problem remained present until solved. However, because of the shorter between-problem interval in Study V, more problems were presented than in Studies I through IV. Measures were recorded separately for first and second solutions as follows: response time per subject measured from the time of occurrence of the last preceding response; number of errors for each subject; and number of correct responses for each subject. The measures used in the analyses were time per response (whether right or wrong) for both first and second solutions separately and the percentage of errors on the second solution.

Task Combinations. In Studies I through IV, the 1-hour test session was broken down into the following 15-minute performance periods (Table 1). For the first 15 minutes, the monitoring tasks (red and green lights and meters) and the mental arithmetic task were active. For the second 15 minutes, the code-lock task was added to these two tasks. During the third 15 minutes, the monitoring task, the pattern discrimination task and the code-lock task were active. And, during the final 15 minutes, the monitoring tasks and the pattern discrimination task were active. In Study V, the first three 15-minute periods involved the same task combinations mentioned above, but the final 15 minutes involved the monitoring tasks plus the code-lock task rather than monitoring plus pattern discrimination.

TABLE 1.—Task Schedule During Performance Testing Studies I–IV

	Time in Minutes			
	0–15	15–30	30–45	45–60
Red Lights	X	X	X	X
Green Lights	X	X	X	X
Meters	X	X	X	X
Arithmetic	X	X		
Problem Solving		X	X	
Pattern Discrimination			X	X

Note: In Study V, problem solving replaced pattern discrimination during the period 45–60.

Training and Testing Procedure. The subjects were seated in open-backed booths that measured 42″x42″x72″. Their chairs were of the standard, armless office type. The center of the MTPB panel was located 38 inches above the floor. Each booth was equipped with a hot-wired, integral earphone microphone intercom system in addition to a hot-wired, desk-type intercom. An 8-inch extended range speaker was located in either the upper right or upper left front corner of the booth; during testing, broadband white noise at approximately 70 to 75 dB (C) was introduced through this speaker to mask out equipment sounds from the control room next door. A half-silvered mirror between the control room and the test room afforded a good view of the open-backed test booths. In addition, a speaker in the control room was wired into both of the subjects' intercom systems.

The subjects were first introduced to the green lights and then the red lights. They were given

several sample signals and told to respond to them as quickly as possible and also to get in the habit of scanning the panel for the occurrence of signals. Next, the meter task was introduced and the method of responding was explained; in Studies I and II, this was a perfunctory procedure; in Studies III, IV, and V, more time was required to explain how to distinguish between the random wandering of the pointers and a signal. The subjects were instructed to respond as quickly as possible on the meter task. However, for Studies III, IV, and V, subjects were cautioned not to make errors, i.e., they were to respond only when they were "reasonably certain" that a signal was present, because false responses would be recorded.

Throughout the remainder of training, signals were introduced on the lights and meters at the same rate as during the test session and the subjects were instructed to respond to them.

The arithmetic task for Studies I, II and V (easy problems) did not require a lot of explanation. The subjects were "talked through" two or three problems and then permitted to work a total of 15 problems. Their progress was observed and any subjects who appeared to be having difficulty were given assistance as required. For Studies III and IV, more time was required. Subjects were "talked through" six problems—two in which "carrying" or "borrowing" were not involved, two in which carrying was involved and two in which borrowing was involved. Subjects in Studies III and IV were given a total of 20 problems during which additional assistance was provided as needed. A blue light flashed as a problem was replaced if the preceding answer had been entered correctly. The subjects were told to work quickly and accurately. With this, and with the remaining tasks, they were reminded to be sure to keep track of the meters and the lights.

For the pattern discrimination task, the subjects were "talked through" two problems and then were permitted to work a total of 10 problems. Additional assistance was seldom required on this task. One of three green lights flashed to tell the subject what the correct answer was on the preceding problem. The subjects were told to respond quickly and accurately.

Problem Solving. This task required perhaps the most explanation of all of the tasks. The subjects were given a description of the task, using much of the information given in the apparatus section of this report. The subjects were "talked through" two or three problems, and, usually, additional assistance was required during the remaining five or six practice problems. For an occasional group, the experimenter had to remind the subjects about some aspect of the task after the test session itself was underway.

Upon completion of the above descriptions, the subjects were reminded how the tasks were scored, that they should be alert to the introduction of meter and light signals at all times, and that they should try to do their best—as though they were actually taking a test that was a prerequisite to their employment as controllers. Approximately two minutes intervened between the completion of these reminders and the start of the test session. The entire instruction and practice period lasted about 50 minutes.

In Studies I through IV, the second and third sessions began without any preliminaries other than to remind the subjects to do their best. When the final testing was completed for a given group, if time permitted, the subjects were allowed to see the programming and scoring equipment and were shown some of the absolute values of their scores; they were given no information about their relative standing with respect to the other groups that had been tested. The subjects in Study V were told that several weeks after all of the testing had been completed they could obtain information about their performance on the MTPB as regards their rank order position in the group of 89 subjects tested.

Criteria

Criterion of Trainee Potential. In Study I, which was strictly an exploratory investigation, grades from the FAA Academy 9-week En Route training course and instructor ratings of each student were both used as criteria. Since these subjects were tested during the time that the course was in progress, current grades were available. The instructor ratings were based on a 3-point scale, with 1 being indicative of a trainee with good potential and 3 being a trainee with marginal potential.

The subjects tested in Studies II and III had completed the 9-week En Route course from three months to a year prior to the time we tested them. Thus, although summary, end-of-

course grades comparable to those for Study I were available, they were not current. In any event, the results of Brokaw's and Trites' investigations suggest that instructor evaluations would be the best available criterion of controller potential. Therefore, our primary interest was in the ratings provided by the instructors from the 4-week interim course which the subjects were attending at the time they were tested. Each of three instructors was asked to provide a rating on each of the trainees in his class who participated in the study. The form used is reproduced in Appendix II. The instructor was asked to indicate his opinion of the trainee's "potential for becoming a fully rated air traffic controller." A rating of 70 was to be considered as indicating below average but acceptable potential; a rating of 90 indicated above average potential. The instructors were told to give their own opinions independent of those of any other instructor; the form administrator remained with the instructors to assure that this caution was observed.

The same procedure for securing instructor ratings was followed in Studies IV and V (which involved trainees from the Academy 9-week Terminal training course). In addition, a full set of grades was available for all phases of the course as well as the end-of-course grades.

Performance Measures Used as Predictor Variables. In Study I, two approaches were used in the construction of the predictor variables. The first approach involved a simple ranking procedure. The 19 subjects were ranked on each of the task measures. The ranks for each of the measures were summed to provide a sum of ranks for each subject; then the subjects were ranked in terms of this sum of ranks. The second approach involved correlational analyses of the individual performance measures against a composite score involving Academy grades and instructor evaluations. However, with only 19 subjects, a correlation of .456 is required for significance at the .05 level. Thus, only very strong relationships between performance measures and the Academy criteria would be detected, i.e., show statistical significance.

Starting with Study II, composite performance indices were used as the predictor variable. The composite consisted of a summation of performance scores, with each individual measure being weighted by the reciprocal of the standard deviation for that measure; the standard deviation used was based on all of the subjects in that particular study. All time measures were subtracted from a constant so that good performance would be associated with a high score for all measures. The specific measures used in the composite were chosen primarily on the basis of our judgment as to the likely relevance of the measure to controller potential. This was moderated by the appearance of consistently low criterion correlations for some of the measures. Specifically, the measures from the problem-solving task during the first solution were found not to bear any consistent relation to the instructor ratings. This probably resulted from the chance factors built into the task, namely, the occurrence of errors and, to an extent, the speed of response were influenced by the relation of the correct sequence for a given problem to the standard search sequence. How this might affect a given subject would depend on his position in the standard search sequence. For the most part, an insufficient number of problems was presented in 30 minutes of performance to balance this factor out across subjects. In addition, the interaction among subjects required in the performance of the task could vary from one group to another; this would also tend to obscure any "true" relation to the criterion. (These problems could, of course, be removed, either by making the task strictly an individual task or, preferably, by letting the testing device serve as the other four "subjects.")

III. Results.

Content Validity of MTPB. The Academy instructors judged the work load demands of the MTPB to be quite similar to those placed on the controller (Table 2). However, as shown in Table 3, the level of demand imposed by the MTPB was judged to be greater than the level encountered by the controller on the job. They

TABLE 2.—Instructor Opinions on Similarity of MTPB and Controller Work Loads

Not at all similar	1	2	3	4	5	6	7	Very similar
frequency	0	1	2	1	13	15	8	

TABLE 3.—Instructor Opinions on Percentage of Time Controller Experiences MTPB Level of Work Load

	10	20	30	40	50	60	70	80	90
frequency	12	7	5	4	3	3	5	1	0

also indicated that they felt the MTPB confronted them with a generally more difficult task than the trained controller typically experiences on the job (Table 4).

TABLE 4.—Instructor Opinions on Difficulty of MTPB in Relation to Job of Controller

Very much less	1	2	3	4	5	6	7 Very much more
frequency	1	2	3	9	12	11	2

The instructors generally rated the "behavioral functions measured by the MTPB" as important to the individual's proficiency as a controller (Table 5). The instructors thought that information on the performance of these tasks would be quite useful in making a decision about a marginal trainee (Table 6); 16 of the 40 instructors indicated that the MTPB measures would be "of very great use."

TABLE 5.—Instructor Opionions on Importance of MTPB Behavioral Functions

Very unimportant	1	2	3	4	5	6	7	Very important
frequency	0	1	1	8	7	12	11	

TABLE 6.—Instructor Opinions on Usefulness of MTPB as Trainee Evaluation Device

Very little use	1	2	3	4	5	6	7	Very great use
frequency	0	0	1	2	12	9	16	

As regards their opinions on the value of the individual tasks and task combinations as predictors of the potential of the trainee to become a good controller, there was good agreement that the group problem-solving task would be the best task. Nineteen instructors ranked this task first and no other task was ranked first by more than six instructors. Their reactions to the other individual tasks were somewhat varied (Table 7). On the average, the ordering of the expected predictive value of the tasks was (1) problem solving, (2) arithmetic, (3) pattern discrimination, (4) green lights, (5) red lights, and (6) meters. (Thirty of the forty instructors were tested on the meters after the random background disturbance had been introduced into the pointers.) The task combination in which arithmetic, monitoring, and problem solving were performed simultaneously was, on the average, judged to be the best predictor of controller potential; 17 instructors ranked this combination first. The combination involving problem solving, pattern discrimination, and monitoring was second; arithmetic and monitoring was third; pattern discrimination and monitoring was ranked fourth (Table 8).

TABLE 7.—Instructor Rankings of Tasks as Predictors of Controller Potential

	Rank					
	6	5	4	3	2	1
Red Lights	13	9	0	4	8	6
Green Lights	5	7	9	8	8	2
Meters	5	8	16	7	2	4
Arithmetic	6	7	6	10	6	5
Problem Solving	2	1	3	5	9	19
Pat. Ident	9	6	5	8	10	3

TABLE 8.—Instructor Rankings of Task Combinations as Predictors of Controller Potential

	Rank			
	4	3	2	1
Monitoring, Arithmetic	9	9	14	2
Monitoring, Arithmetic, Problem Solving	4	4	9	17
Monitoring, Pat. Ident., Problem Solving	4	11	9	10
Monitoring, Pat. Ident	17	9	4	4

Characteristics of the Criteria of Controller Potential. The distribution of instructor ratings for the subjects in Study I is shown in Table 9.

TABLE 9.—Frequency Distribution of Instructor Ratings in Study I

	Mean Rating				
	1	1.5 1.6	2	2.5	3
frequency	7	4	4	1	3

In the judgment of the instructors, 11 of the 19 subjects had either good or superior potential of becoming fully-rated controllers; three subjects were rated as poor to marginal in potential. The distributions of instructor ratings for Studies II through V are shown in Table 10.

TABLE 10.—Frequency Distribution of Instructor Ratings

	II	III	IV	V	Total
95–100	7		1		8
90–94	7	10	9	7	33
85–89	23	10	7	13	53
80–84	9	7	7	33	56
75–79	9	4	1	18	32
70–74	4		3	5	12
65–69	1		1	7	9
60–64			0	3	3
55–59			0	2	2
50–54			0	1	1
(15)			1		1
Mean	84.8	85.7	82.9	79.2	
S.D.	6.7	4.8	13.9	8.4	
N	60	31	30	89	

The ratings in all four of these later studies were bunched toward the high end of the distribution. In Study II, 37 out of 60 of the subjects were rated above the mean. In Study III, 18 of the 31 subjects were given ratings above the mean. In Study IV, 19 of the 30 subjects were given ratings above the mean. However, if one subject with a mean rating of 15 is excluded from Study IV, the mean rating for the group becomes 85.1; for the remaining group of 29 subjects, 16 were given ratings above the mean. In Study V, 53 of the 89 subjects were given ratings that exceeded the mean for the group.

There was one trainee in Study I who failed the Academy training course. A second trainee withdrew voluntarily, reportedly because he expected to be unable to pass the remainder of the course. The subjects in Studies II and III were attending an advanced, ungraded course in which there would be no failures. One trainee in Study IV failed; this subject was said by other trainees to have failed simply because he did not apply himself to the task of studying. Ten trainees in Study V failed and one withdrew voluntarily because of a family illness. Three other subjects, although they passed the course, were judged by the instructors to have unacceptable potential for becoming fully-rated controllers.

The instructor ratings for the subjects of Study V were further analyzed to obtain estimates of the reliabilities of the ratings. Intra-class correlation coefficients were computed for each sub-set of subjects who were rated by a given set of three instructors. Four of these sub-sets contained 20 subjects and one contained nine subjects. The intra-class correlations ranged from .52 to .74 (Table 11). These correlations can be interpreted as being an estimate of the reliability of the ratings given by a single instructor. The computed correlations for the mean of three ratings ranged from .77 to .89; these correlations estimate the reliability of the average of three instructor ratings for the indicated number of trainees. These latter correlations were normalized by conversion to inverse hyperbolic tangents; the inverse hyperbolic tangents were averaged and then converted back to an average correlation. This average correlation was .84; it provides an estimate of the reliability of the instructor ratings for the entire group of 89 trainees.

TABLE 11.—Intra-class Correlation of Instructor Ratings for Study V

Class Designation	N	Intra-class r	Predicted Reliability of Mean
15	20	.74	.89
16	20	.57	.80
17	20	.52	.77
18	20	.68	.87
19	9	.59	.81

The intercorrelation matrix of Academy grades and the instructor ratings is shown in Table 12. The pre-phase grade was the score on a non-credit examination given at the beginning of the Terminal training course. The next five grades involve written examinations covering the subject matter implied by the name. With the exception of the "local control grade," there was a corresponding grade assigned on the basis of performance of a simulated task in the laboratory. There was also a final written examination given at the end of the course. The instructor rating appears to depend most heavily on the trainee's performance in the non-radar laboratory test (r=.72).

A multiple regression analysis of the Academy grades as predictors of the instructor ratings yielded a multiple correlation of .84 using only the non-radar laboratory, radar laboratory, and flight data laboratory grades; the partial correlations of the remaining grades with the instructor ratings at this point in the analysis fell well below the value required for significance at the .05 level.

Grades were also obtained on the 19 subjects of Study I. These were the final laboratory examination (referred to at the Academy as "D position final"); the average of all written tests during the course; the grade from the strip-writing laboratory; the grade from the flight data laboratory; and the final written exam grade. Three of the correlations between the instructor ratings and these grades were significant—the D position final (r=+.92); the flight data laboratory (r=+.70); and the final written exam (r=+.61).

End-of-course grades were available on 59 of the subjects of Study II. For these subjects, the grades had been assigned during a previous course which ended from three months to more than a year prior to the time at which the current instructor ratings were obtained. In most cases, different instructors were involved in the assignment of the ratings and the assignment of the previous grades; the instructors did not have access to those grades at the time the ratings were made. The correlations among these measures were as follows: final laboratory examination vs. instructor rating, r=+.46; final written examination vs. instructor rating, r=+.47; final laboratory examination vs. final written examination, r=+.32.

Multiple Regression Analyses. At the inception of this research effort, multiple regression analysis of the MTPB measures and the criterion seemed like the most reasonable approach to follow. However, we were aware of the fact that the required assumption about the independence of the MTPB measures was tenuous at best. (A multiple regression analysis was performed on the data of Study I, but, with only 19 subjects, this had to be regarded as only an exercise.)

The second study, with a total of 60 subjects, provided a minimal condition for application of the multiple regression analysis technique. But, as in Study I, it was necessary to ignore the

TABLE 12.—Intercorrelation Matrix of Academy Scores and Performance Composite for Study V

	PP	FD	GC	LC	NR	R	FDL	GCL	MRL	RL	FINAL	IR	PC
Prephase	—	.30	.29	.28	.26	.53	.28	.32	.33	.32	.73	.45	.35
Flight Data		—	.18	.25	.40	.26	.24	.31	.34	.25	.26	.34	.09
Ground Control			—	.26	.41	.35	.24	.18	.06	.18	.35	.27	.01
Local Control				—	.40	.50	.09	.36	.34	.30	.44	.40	.12
Non Radar					—	.35	.31	.26	.39	.32	.44	.43	.11
Radar						—	.34	.32	.31	.23	.73	.44	.18
Flight Data Lab							—	.36	.32	.29	.37	.52	.47
Ground Control Lab								—	.40	.31	.36	.55	.19
Non Radar Lab									—	.21	.27	.72	.30
Radar Lab										—	.32	.50	.25
Final											—	.48	.07
Instructor Rating												—	.46
Performance Composite													—

N=76 through 89 For N=76 $r_{.05}$=.23 $r_{.01}$=.31

usual rules of significance of the individual validities to proceed. Since we wanted to examine, at least approximately, the nature of results one might expect from this approach, we proceeded. Thus, the actual results of the analyses are relevant only to methodology. The total sample was divided into two groups, one containing 28 and the other 32 subjects. The data of each of the three test sessions were then subjected to a multiple regression analysis using a computer. The regression equation for each sub-sample was then used to predict the criterion for the other sub-sample. The criterion of primary interest was the instructor ratings of the trainee's potential. However, two other criteria were also used. The first of these combined the instructor ratings with two of the end-of-course grades from the earlier Academy course; the instructor rating was given a weight of .2985; the final laboratory grade was given a weight of .3656; and the final written exam was given a weight of .2452. The second of these additional criteria combined the instructor rating (a weight of .2633), the final laboratory grade (a weight of .3777), the final written exam (a weight of .2783), and the strip writing laboratory grade (a weight of .1495). For both of these additional criteria, the weights were those derived in Study I.

The results of these analyses are shown in Table 13. Of the total of 18 multiple correlation coefficients computed, five were significant at the .05 level of confidence. In only one case did the R's for both sub-groups for any combination of criterion and test session reach significance, namely, criterion 2, session 2. None of the cross-validation correlations associated with these significant multiple R's was significant. However, three of the 18 cross-validation r's, though not associated with significant multiple R's, were significant, and two of these matched, i.e., the same criterion and session were involved with the regression equation for the first group predicting the criterion for the second group and vice versa. It was because of the generally confusing picture presented by these correlations that we switched to a performance composite that simply gave approximately equal weight to each of the MTPB measures. Apparently, either our data did not fit the multiple regression analysis paradigm, or the MTPB measures did not have reliable predictive power for the criteria.

The Equal-Weight Composite of MTPB Measures. The data from Study II were re-analyzed using the performance composite computed as described in the Method section of this report. Ten measures were used, all being taken from the second and third 15-minute performance periods. They were: separate response-time measures for red lights, green lights, meters, and arithmetic; percentage correct for arithmetic and pattern discrimination; and separate time-per-response measures for the group problem-solving task for first and second solutions, both with concurrent performance of arithmetic and concurrent performance of pattern discrimina-

TABLE 13.—Results of Multiple Regression Analysis

	Criterion 1		Criterion 2		Criterion 3	
	Obtained R	Cross Validation r	Obtained R	Cross Validation r	Obtained R	Cross Validation r
Session I						
Group I N=29	.5970	.2861	.6763	.4111*	.6867	.2099
Group II N=31	.6473	.2279	.6988	.4108*	.6781	.0057
Session II						
Group I N=29	.6737*	.1682	.7958*	.3419	.6370	.2026
Group II N=31	.5970	.2163	.6930*	.1869	.7108	.4462*
Session III						
Group I N=29	.7119	−.1386	.6361	−.1781	.7955	.0230
Group II N=32	.7321*	.06	.7577*	.0513	.7052	.0875

*r significant at .05 Note: The number of predictor variables used differed across analyses.

tion (a total of four problem-solving measures were used).

The session-to-session reliabilities of the performance composite were: session 1 versus session 2, .87; session 1 versus session 3, .73; and session 2 versus session 3, .73. All of these test-retest reliabilities of the MTPB composite are significant at the .01 level of confidence.

The correlations of this composite for each session and each criterion are shown in Table 14.

TABLE 14.—Correlations with Performance Composite

	N	Session I	Session II	Session III
Criterion 1				
All Subjects	61	.3537**	.2797*	.2686*
Ss with CSC	28	.4089*	.3031	.2358
Ss without CSC	31	.2322	.2111	.2386
Criterion 2				
All Subjects	59	.4101**	.3196*	.2966*
Ss with CSC	28	.3829*	.2776	.2461
Ss without CSC	31	.2586	.2399	.2073
Criterion 3				
All Subjects	61	.5027**	.4203**	.3460**
Ss with CSC	28	.3126	.2260	.1595
Ss without CSC	32	.5900**	.4958**	.4366**

* $p < .05$
** $p < .01$

Three points stand out as regards these data. First, all of the correlations are significant at the .05 level of confidence or better. Second, the highest correlation with each criterion was found on session 1. And, third, the highest correlation for each session was found for the instructor rating. An optimized criterion derived by multiple regression analysis of the Academy grades and instructor ratings against the performance composite increased the correlation only slightly (R=.57).

The correlation matrix of the 10 MTPB measures for session 1 raw data and the instructor ratings is shown in Table 15. The data are based on the entire Study-II sample (N=60). Five of the correlations with the criterion are significant at the .05 level or better; a sixth measure is just short of significance. All six of these correlations involve measures of response time. It is clear from this table that a composite based on all 10 measures includes measures that would be expected to contribute little to the resultant validity for the criterion. In particular, the first-solution measures for the problem-solving task stand out as non-predictive. And the correlations with the arithmetic and pattern discrimination tasks, though larger, are not significant and do not make much of a contribution. Subsequent to these analyses, another measure was examined, namely, the percentage of correct responses on the second solution for the problem-solving task (both with arithmetic and with pattern discrimination). This measure correlated +.34 with the criterion ($p<.01$). A new composite was then constructed in which the first-solution measures on problem solving were dropped and percentage of correct responses on the second solution was added. The correlation between the instructor rating and this new composite was .54.

The pattern of decreasing correlations from the first to the third sessions is what led us to change the task structures. Our reasoning was as follows: The decrease in correlation might have been a result of a factor analogous to a dynamic change in item difficulty. As learning progressed and the tasks became easier for the subjects, less discriminative power would be available if the "item" difficulty in effect became too easy. Since we would expect the MTPB measures to be more stable later in learning, it was possible that, by increasing the overall difficulty of the task complex, we would achieve increased reliability and predictive power on session 2 or session 3 relative to session 1. Therefore, the random background disturbance was introduced into the meters and more difficult arithmetic problems were used in Studies III and IV.

The results of Study III did not confirm this hypothesis about task difficulty. The correlation with the instructor rating on session 1 was .38 for the original, 10-measure composite ($p<.05$) and .53 for the revised composite ($p<.01$). The correlation for the second session with the original, 10-measure composite was .12; for the third session it was .14. Apparently, whatever it is that the MTPB measures that has predictive power for the instructor ratings is measured best on the first session.

Unfortunately, Study IV got underway before the results of Study III were completely analyzed. Thus, the same task conditions were

TABLE 15.—Intercorrelation Matrix of Performance Measures and Instructor Rating, Study II

	R	G	M	At	A%	P%	S15	C15	S30	C30	IR
Red	1.0	.24	.16	.29	−.17	−.33	.05	.22	.12	.21	−.38
Green		1.0	.38	.33	−.31	−.32	−.19	.07	−.11	.12	−.27
Meters			1.0	.39	−.12	−.30	−.02	.18	.11	.36	−.34
Arithmetic time				1.0	−.59	−.46	.22	.04	.18	.26	−.25
Arithmetic %					1.0	.36	.03	.09	−.15	−.09	.14
Pattern Discrimination %						1.0	.06	−.09	.11	−.13	.17
Problem Solving, Solution Phase, 15							1.0	.18	.20	−.06	−.01
Problem Solving, Confirmation Phase, 15								1.0	.06	.22	−.26
Problem Solving, Solution Phase, 30									1.0	.26	−.09
Problem Solving, Confirmation Phase, 30										1.0	−.42
Instructor Rating											1.0

N=60
$r_{.05}$=.25 $r_{.01}$=.33

used as with Study III. In this case, a significant *negative* correlation was found on session 1 between the 10-measure composite and the instructor rating ($r=-.40$, $p<.05$). The subject who scored the highest on the MTPB composite was the only subject who failed the Academy course; his instructor rating was 50 units below that of the next lowest subject. (This was the subject who was reported to have not applied himself to his studies.) The correlation without that subject included was +.07. The correlations on sessions 2 and 3 were also essentially zero. With this sample of subjects and this task configuration (if the legitimacy of dropping the one aberrant subject is allowed), there was no relation between MTPB performance and the instructor ratings of trainee potential.

In Study V, the correlation between the instructor rating and the revised performance composite for the single test session was .24 ($p<.05$). It could be argued that, because the task situation was changed in Study V, it should not be surprising that a performance composite computed in exactly the same way as in the earlier studies did not correlate as well with the instructor ratings. As was pointed out in the Procedure section of this report, the signal rate on the red and green lights was approximately 106 signals in the one-hour test session for the two tasks combined. This is an increase of approximately 50% over the rate used in Studies I through IV. The signal rate on the meter task was approximately 110 signals per hour; this was substantially higher than in Studies III and IV but the same as in Studies I and II; however, in Studies I and II there was no random function causing the pointers to wander about. The inter-solution interval on the problem-solving task, which was 30 seconds in the earlier studies, was reduced to 15 seconds in Study V. The greatest effect of these changes would be expected to occur during the two heavy workload periods, i.e., when arithmetic and pattern discrimination were being performed. A different composite was constructed to see if a higher correlation would be obtained if the data used were from only those periods of lighter work load, namely, the first 15 minutes when only arithmetic and monitoring were performed and the last 15 minutes when only problem solving and monitoring were performed. This new composite, which should have been affected less by the task changes, consisted of all the monitoring data from the first and last 15 minutes plus the problem-solving data from the last 15 minutes. The correlation between this composite and the instructor rating was +.41 ($p<.01$). When the problem solving and the red and green lights monitoring data from the second and third periods and the pattern discrimination data were added to the above-described composite, the correlation increased slightly ($r=+.46$). The correlations between this last-mentioned composite and the course grades given by the Academy are shown in the last column of Table 12. Significant correlations were found for three of the four laboratory grades; only the ground control laboratory grade failed to correlate significantly with the performance composite. A significant correlation was also found between the perform-

TABLE 16.—Correlations of Individual Measures with the Instructor Ratings

	Study Number					
	II N=59	III N=31	IV N=29	V N=89		
Red	.37*	.22	−.25	.28***	.21	.18
Green	.27*	.45*	.02	.20	.31***	.18
Meters	.34*	.30	−.08	.04	.17	.11
Arithmetic time	.24	−.10	−.11	−.05	.12	
Arithmetic %	.14	.28	.06	.04	−.02	
Pattern Discrimination %	.17	.41*	.09	.26**		
Problem Solving, Sol 15	.01	−.19	−.11	.20		
Problem Solving, Con 15	.26*	.12	.04	.19		
Problem Solving, Sol 30	.09	−.21	−.17	.21		
Problem Solving, Con 30	.42*	.30	.01	.10		
Problem Solving, Con % 15+30	.34*	.20		.15		
Problem Solving, Con % 45						.16
Problem Solving, Sol 45						.19
Problem Solving, Con 45						.38***

N=60, r $_{.05}$=.26; N=89, r $_{.05}$= .22, r $_{.02}$= .25, r $_{.01}$= .28; N=30, r $_{.05}$=.36

* p <.05 ** p <.02 *** p <.001

ance composite and the "pre-phase" written examination, but none of the correlations with the other written examinations was significant.

Validities of Individual Performance Measures. The correlations between the individual performance measures and the instructor ratings are shown in Table 16. Because complete session-1 data were available on only 10 subjects in Study I, those correlations are not included.

The measure of response time to the onset of red lights correlated significantly with the instructor ratings in Studies II and V. Response time to the offset of green lights correlated significantly in Studies II, III, and V. Response time to the meter task correlated significantly with the instructor ratings only in Study II, the study in which the pointers were stationary except for the occurrence of a signal. The time required to solve the arithmetic problems yielded a correlation that approached significance in Study II but the correlations were quite small in the other studies. The measure of percentage correct on the arithmetic task was not significantly related to the instructor ratings in any study. Percentage correct on pattern discrimination correlated significantly with the instructor ratings in Studies III and V. The data for the first solution on the group problem-solving task did not correlate significantly with the instructor ratings in any of the studies. The time-per-response for the second solution correlated significantly with the instructor rating when problem solving occurred both with arithmetic and with pattern discrimination in Study II. This measure also yielded a significant correlation in Study V during the final 15-minute period when problem solving was performed with only the monitoring tasks. The proportion of errors on the second solution in the group problem-solving task correlated significantly with the instructor ratings only in Study II; the data from both the second and third 15-minute periods were combined for this measure.

Correlations with the CSC Battery. The Composite Score from the Civil Service Commission Air Traffic Control Specialist Screening Battery was obtained for 28 subjects in Study II and for 60 subjects in Study V. In addition, CSC scores and performance data were available from another study on 145 journeyman controllers (83 Terminal and 62 En Route).* The intercorrelations of these scores with the performance measures and, for Studies II and

* The CSC scores were obtained under Contract DOT-FA70WA-2371 between the FAA and Education and Public Affairs, Inc.

V, with the instructor ratings, are shown in Table 17.

TABLE 17.—Intercorrelations with the CSC Battery Composite

		r
	Study II	
N=28	CSC vs. IR	.26
	Perf vs. IR	.31
	CSC vs. Perf	.23
N=32	*Perf vs. IR	.59
	Study V	
N=60	CSC vs. IR	.40
	Perf vs. IR	.36
	CSC vs. Perf	.32
N=29	*Perf vs. IR	.47
N=60	CSC vs. Crit**	.39
	Perf vs. Crit	.39
N=29	*Perf vs. Crit	.46

*Subjects for whom CSC scores were not available.
**Composite criterion: 2 IR + 2 Lab ave. and written exam ave.

None of the intercorrelations in Study II were significant. However, because of the small N, a correlation of .37 is required for significance at the .05 level of confidence. The correlation between the performance composite and the instructor rating for the 32 subjects for whom we did not have CSC scores was substantial ($r=+.59$; $p<.01$). All three of the intercorrelations in Study V were significant at the .02 level of confidence or better. The correlation between the performance composite and the instructor ratings for the 29 subjects in Study V for whom we did not have CSC scores was +.47 ($p<.01$). For the purpose of this analysis, we also constructed another criterion for the subjects of Study V; it was a composite of the instructor ratings (given a weight of 2), the average Academy laboratory grade (given a weight of 2), and the average Academy examination grade (given a weight of 1). The correlation between the CSC scores and this criterion was +.39; the correlation between the performance composite and this criterion was also +.39 ($p<.01$ in both cases). The correlation between a very similar performance composite and scores on the CSC Battery for the 145 journeymen (who took the test experimentally for the purposes of another study) was +.46 ($p<.01$).*

Effects of Manipulations of Task Difficulty. Inspection of the correlations shown in Table 16 suggests that the increase in signal rate on the red and green lights in Study V did not alter the relation of those measures to the instructor ratings. Introduction of the random background disturbance into the meters (Studies III, IV, and V) appeared to decrease the predictive power of the response-time measure on that task, as evidenced by the lower correlations found in these later studies. The increase in the difficulty of the arithmetic problems in Studies III and IV resulted in much smaller correlations between the arithmetic response-time measure and the instructor ratings than were found in Study II. (The correlations with the percentage-correct measure on arithmetic were not significant in any of the studies.) Although the arithmetic problems used in Study V were the same (easy) problems that were used in Study II, the correlation for arithmetic time was negligible, perhaps because of the increased rate of presentation of problems on the concurrently performed group problem-solving task. As far as the problem-solving task itself is concerned, the increase in rate of presentation of problems in Study V did not improve the correlations for that task. In fact, it worsened the situation; the only significant correlation on that task was found during the final 15-minute period when problem solving was performed by itself.

The first-session means and standard deviations are shown in Table 18 for each measure for each study. The results of an analysis of variance is also shown; in each case the analysis did not include the data from Study I. For only one measure, the second solution on problem solving when performed concurrently with arithmetic, was the F for the differences across studies not significant. For the response times to red lights the F was significant at the .01 level of confidence, and, for all other measures, it was significant at better than the .001 level of confidence. Table 19 summarizes the results of "t" tests applied to the individual means. The domi-

* The correlations between a very similar MTPB composite measure and the individual CSC battery tests are shown in Appendix III along with correlations with a number of other written tests.

TABLE 18.—First Session Means and Standard Deviations

Session I		Red	Green	Met	At	A%	P%	S_{15}	C_{15}	S_{30}	C_{30}
Group I											
N=10	M	1.96	6.74	4.19	7.62	78.11	71.26	2.57	2.03	1.75	1.61
	SD	.52	5.12	1.66	1.12	10.52	28.51	.73	.65	.45	.46
	SD^2	.27	26.18	2.76	1.25	110.78	812.76	.53	.42	.21	.22
Group II											
N=60	M	2.82	10.50	6.14	8.72	73.77	64.09	2.38	2.03	2.05	1.63
	SD	1.52	4.20	2.07	1.58	13.24	14.85	.66	1.60	.57	.58
	SD^2	2.30	17.67	4.28	2.51	175.31	220.49	.43	2.55	.33	.34
Group III											
N=31	M	3.21	11.09	15.52	11.26	56.29	61.07	2.90	2.54	2.06	2.09
	SD	2.17	4.53	6.08	1.06	18.02	17.42	.98	.96	.69	.83
	SD^2	4.71	20.52	36.97	1.12	324.72	303.46	.96	.92	.48	.69
Group IV											
N=29	M	3.88	12.09	23.57	11.56	49.11	67.86	2.88	2.29	2.41	2.34
	SD	1.87	4.25	10.35	1.26	21.51	21.72	1.09	1.02	.96	1.04
	SD^2	3.50	18.05	107.08	1.60	462.70	471.74	1.19	1.04	.92	1.08
Group V											
N=89	M	2.56	7.34	18.92	10.32	53.96	51.06	3.13	2.42	2.59	2.03
	SD	1.37	2.60	14.42	1.63	22.54	20.70	.42	1.52	.79	.90
	SD^2	1.88	6.77	207.88	2.65	508.24	428.53	.18	2.31	.62	.82
F (d.f.=3,206)		5.35	18.80	22.73	32.30	15.69	8.83	13.08	1.13	7.56	5.55
P<		.01	.001	.001	.001	.001	.001	.001	NS	.001	.001

ant factor in Table 19 appears to be the differnces between the performance of the subjects f Study II and the performance of the subjects f the remaining studies. Only in the case of ıe green lights response-time measure for Study r was the performance of any group superior) that of the subjects of Study II. The subcts of Studies III and IV performed signifiıntly differently on only one task; the Study-III subjects were significantly better on the meters. Since there were no changes in the tasks or procedures from Study III to Study IV, this affords a relatively clean comparison of the entering Terminal trainees with the En Route trainees who were well into their training program. In the comparisons of Study III versus Study V, the Study-III subjects performed better on pattern discrimination and on problem

TABLE 19.—Comparison of Mean Scores Across Studies II through V

Study (a) vs. Study (b)	(a) Better	No Difference	(b) Better
vs. III	M, At, A%, S15, C30	R, G, P%, C15, S30	
vs. IV	R, M, At, A%, S15. S30, C30	G, P%, C15	
vs. V	M, At, A%, P%, S15, S30, C30	R, C15	G
I vs. IV	M	R, G, At, A%, P%, S15, S30, C15, C30	
I vs. V	P%, S30	M, A%, S15, C15, C30	R, G, At
vs. V	P%	A%, S15, C15, C30, S30	R, G, M, At

ey to Abbreviations

—Red lights
—Green lights
—Meters
—Arithmetic time

A%—Arithmetic % correct
P%—Pattern discrimination % correct
S15, S30—Problem solving, solution phase, 15 or 30 minute interval
C15, C30—Problem solving, confirmation phase, 15 or 30 minute interval

solving (time-per-response, first solution) when those two tasks were performed concurrently. The Study-V subjects performed better than the Study-III subjects on red and green lights and they solved the arithmetic problems (which were easier) more quickly. The Study-IV subjects performed better than those of Study V on the pattern-discrimination task, but the Study-V subjects performed better on the red and green lights, on the meters, and they solved the arithmetic problems more quickly.

The greater difficulty of the arithmetic problems presented to the subjects in Studies III and IV was reflected in the significant differences between each of these two studies and the performance of the Study-II subjects, both in terms of percentage correct and in terms of response times. The Study-V subjects, who also worked the easier arithmetic problems, were faster but no more accurate in arithmetic performance than were the subjects of Studies III and IV.

Introduction of the random background disturbance into the meters had a very large, consistent effect on the performance of that task, as is readily seen when Study II is compared to the other studies. The increased rate of presentation of signals on the red and green lights *appears* to have resulted in better performance at the higher signal rate (Study V versus the other studies). However, other differences in the task structures across studies are confounded with this effect.

The decrease in the inter-problem interval on the group problem-solving task (and this effectively increased the rate of presentation of problems) appears to have resulted in poorer performance of this task. The performance of the subjects in Study II was superior to that of the subjects in Study V on three of the four problem-solving measures.

IV. Discussion.

Instructor Ratings of Students and the MTPB. A number of FAA Academy instructors were involved in these studies and they played important roles in two respects—as expert evaluators of the potential of the MTPB as a selection device and as raters of the trainees' potentials for becoming fully-rated controllers. The qualifications of these instructors to perform these two functions are pivotal to the evaluation of the results of these two aspects of the studies.

Two major qualifications for being selected as an instructor at the Academy are relevant to this discussion. First, the instructor must be a qualified journeyman controller, and, second, he must have been given "better than average" ratings on his Employee Appraisal Records.

The Academy controller training programs are directed at instilling in the trainee those skills and knowledges necessary to becoming a qualified controller. Thus, the instructor is reminded on a day-to-day basis of what are considered to be the essential components of the controller's job. The laboratory facilities at the Academy would appear to lend themselves to the development in the instructor of a good understanding of these factors. The instructor's experience in the laboratory should also provide him with a good basis for understanding the potential of trainees for becoming fully-rated controllers. Here, the instructors have an opportunity to observe the trainee under reasonably well-controlled conditions. Over a period of time, the instructor gets to observe many different trainees performing in the same situation on the same simulated problems. These observations should permit him to develop a good idea of the range of skills exhibited by trainees in performing controller tasks, and the observations should also permit him to develop a good grasp of the kinds of abilities that are essential to good training performance and, by inference, to becoming a proficient controller.

For these reasons, we would contend that both the opinions of the instructors with respect to the potential of the MTPB and their ratings of trainee potential command respect. In this latter instance, the findings reported by Brokaw (1959) and by Trites (1961) on the validities of instructor ratings as predictors of later supervisory ratings lend strong support to this stand. In the former instance, the correlations we found between our performance measures and the instructor ratings lend at least indirect support to the contention that the instructor evaluations of the selection potential of the MTPB are reasonably reliable.

The overall reactions of the instructors to the MTPB, as reflected in the evaluation questionnaires, were quite positive. We interpret the findings to indicate that the instructors think the measures from the MTPB ought to be useful

predictors of the likelihood that a trainee would n fact become a fully-rated controller.

The accuracy with which the instructors anked the tasks with respect to their expected predictive efficiency should probably not be udged too severely. The instructors readily ccepted the notion that the predictive power of he MTPB would most likely derive from its bility to measure time-sharing skills. From his point of view, the differential validities of he individual tasks would not be considered to e very important.

The shape of the distributions of instructor atings reflects the nature of the procedure used n selecting applicants to enter the controller raining program. Specifically, applicants are elected from the register of those qualified by tarting with those most qualified and working own until the requirements for new personnel re met. This procedure would be expected to esult in some degree of negative skewness. Considering the subjects in Studies II through as a whole, 70% of the trainees were judged y the instructors to have average or better potntial for becoming fully-qualified controllers. The results of the selection procedure can also e seen in the high average CSC scores of the bjects for whom such scores were available. n Study II, the 28 subjects for whom we had cores had a mean composite CSC score of 247. n Study V, the 60 subjects for whom we had cores had a mean composite CSC score of 242.

The estimated reliability of .84 found for the structor ratings in Study V was encouragingly gh. However, it is clear that the interactions mong the instructors in the conduct of the urse would lead to some degree of interpendence among the ratings from any given t of three instructors. Obviously, they would ve discussed among themselves the strengths d weaknesses of the individual trainees as the urse progressed. They would also have interted with each other in developing the infortion to be used as a part of the official cademy evaluation of the trainee's performance. us, the estimated reliability of the instructor tings is probably somewhat inflated, despite the t that care was taken to assure the independence of the three ratings per trainee in terms of e actual elicitation of the ratings from the inuctors. The amount of the inflation is unown.

Academy Training Performance and the Instructor Ratings. It did not seem reasonable to use a pass/fail criterion to evaluate the predictive power of the MTPB. Studies I, IV, and V were the only studies in which subjects actually failed. The subjects of Study I should probably not be considered because they were given a briefing that did not mention selection as being the intent of the research. The small number of failures in Studies IV and V and the fact that the two studies involved different task structures militate against a meaningful pass/fail analysis. The failure rate in the Terminal program, prior to the conduct of Studies IV and V, had been running between 20 and 30% according to Academy personnel. However, at about the time these studies were being planned, it became known that the Terminal training program of the Academy was to undergo a significant change. The aspect of the change that has relevance to the failure rate was the fact that, in the new program, trainees would not be eliminated by the Academy; removal from the program would take place after completion of the course and return to the home facility. It is fair to say that the instructors were somewhat more reluctant to eliminate a trainee in view of the fact that trainees in up-coming classes would not be subject to failure. This was the likely reason for the reduced failure rate and for the fact that three of the Study-V trainees were passed even though they were rated as having less than acceptable potential of becoming fully-rated controllers.

Terminal training instructors and supervisory personnel had indicated in previous discussions that it was their opinion that the best predictor (from the Terminal program) of the trainee's potential was his performance in the non-radar laboratory. The intercorrelation matrix of Academy grades and the instructor ratings for Study-V subjects indicates that the instructors do in fact place heavy emphasis on performance in the non-radar laboratory in making their evaluations of trainee potential. The grades assigned in this phase of training plus the grades from the radar and the flight data laboratories account for more than 70% of the variance in the instructor ratings in a multiple regression analysis in which instructor rating is the dependent variable. The instructors apparently gave heavy weight to their recollections of the

trainee's performance in the laboratory exercises when they assigned the ratings.

The ratings assigned by the instructors in Study I were highly related to the grades assigned in the D position final exam; the exam grade can be used to account for almost 85% of the variance of the instructor ratings. In Study II, although the grades from the previous course were assigned by different instructors and were not available to the instructors who rated the Study-II subjects, there were significant correlations between the laboratory grades and the final written exam grade. Thus, the instructors, to some extent, were apparently reacting to the same kinds of behavior in assigning ratings that the original instructors had considered in assigning course grades. Presumably, the current ratings assigned to the trainees were heavily weighted by the subjects' laboratory performance in the advanced course since that was the aspect of the course that was given primary emphasis.

In summarizing this information concerning the nature and apparent sources of the instructor ratings of the trainee's potential for becoming a fully-rated controller, it can be said that these ratings are based largely on the trainee's performance in the laboratory exercises. Other factors enter into the assignment of ratings in some undetermined way, but the contribution of these other factors is apparently not very great.

Multiple Regression Analysis. It came as no great surprise to us that the mutliple regression analyses did not yield regression equations that would cross-validate. The key to good performance on the MTPB is the ability to shift rapidly from attending to one kind of activity or process to another and to quickly adopt the proper set for the new activity. In this study, as in previous research with the MTPB, no attempt was made to structure the tasks in terms of importance. The emphasis was placed on performing all tasks well. Especially when the subjects are first acquiring skill on the individual tasks and in time-sharing the tasks, it would be unrealistic to assume that the subjects assign equal priorities to all tasks. It would be equally unrealistic to assume that different subjects assign the same priority to a given task. Because it is readily apparent to the other subjects that an individual is not doing well on the group problem-solving task, there is a substantial amount of "social" pressure on the subject to place a high priority on problem solving. (Despite that fact, it was not infrequently the case that a subject had to be reminded repeatedly by the other subjects that it was his turn to push his problem-solving button or, conversely, that he had pushed out of turn.) Some subjects may have given the highest priority to the task or tasks they were good at; others may have given priority to tasks they were poor at; and still others may have given priority to the tasks they felt to be most relevant to the job of a controller. Other than the possible exception of the group problem-solving task, defensible speculations as to the priorities assigned to the different tasks can not be offered

The foregoing discussion relates to the appropriateness of the multiple regression analysis technique in the following manner. The validity of a measure from a given task, as indicated by the correlation of that measure with the criterion depends only in part on the relation to the criterion of the *skill dimension* underlying performance of that task. The amount of their available "channel capacity" that the subject devote to the performance of that task and their skill at time sharing will clearly make very important contributions to the apparent validity of the measure. The regression equation resulting from a multiple regression analysis assigns differential weights to the different measures with multiple task performance these weights are to an important extent reflective of the priorities assigned to the different tasks by the subjects. There is clearly no reason to assume that subjects who were given instructor rating within some specific range of values would assign the same priorities to the tasks or would even assign priorities to the tasks on the same basis. Only those subjects who could readily handle the work load of the task complex as a whole would have been expected to assign priorities with any degree of uniformity. Thus, to the extent that the priority assigned will in fact determine the level of performance on a given task, there is an unpredictable element in the situation that contra-indicates the use of regression weights derived from a specific subject sample. Quite possibly, if the sample were large enough, the effects of priority assignment might tend to balance out. But this is not likely to happen with sub-groups of size 30 or 40. F

these reasons, we are inclined to regard the obtained three instances of significant cross-validation correlations as being properly attributed to chance; this interpretation is especially appropriate in view of the fact that the "parent" multiple correlations were not significant in any of those cases.

The Composite Performance Index. The equal-weight composite-performance index evades the problem of specifying differential weights to the various task measures; by using the reciprocal of the standard deviation as the weight, each measure contributes equally in determining the composite score. This approach is somewhat more comfortable in that we do not have any good *a priori* basis for determining differential weights. We could have used the instructor rankings of the tasks as a basis for weighting measures, but the rankings were by no means that clear-cut. A major appeal of the equal-weight composite is that it would seem to give some weight to the ability of the individual to time share the different functions required by the MTPB. Specifically, the subject who concentrated on the performance of one task to the detriment of other tasks was limited as to the extent to which his total performance would be categorized as being good. At a minimum, we could argue that the nature of the MTPB is such that an individual who turns in a very good composite score must be better at time sharing than an individual who turns in a mediocre score. Unfortunately, the concept of time-sharing ability as a unique quantity has not yielded to available mensurational procedures.

The reliability of the performance composite is seen in the correlation between session 1 and session 2 in Study II is quite high. It compares favorably with reliability coefficients from almost any source within the behavioral sciences. The correlation coefficients found in Studies II and III between the performance composite and the instructor ratings are substantially higher than we had any reason to expect at the inception of this program. One does not usually find a correlation as high as .53 between variables of this sort. As Cohen (1969, p. 78) states in his book on the analysis of the power of statistical tests, ". . . when an investigator anticipates a degree of correlation between two different variables 'about as high as they come,' this would by our definition be . . . r=.50." The correlation of .24 found in Study V with the original 10-measure composite is closer to the value we might expect. Cohen (p. 77) describes such a value as being of the magnitude of ". . . many of the correlation coefficients encountered in behavioral science." The interpretation of the correlation of .41 using the data of the first and fourth 15-minute period and the correlation of .46 with the more complex, revised composite depends on the degree of acceptance of the rationale underlying these latter composites. The composites themselves are, of course, quite defensible. The only question concerns the relation of Study V to the earlier studies with respect to the notion of cross-validation. (In any event, let the reader be reminded that in no sense do these studies purport to provide a conclusive, cross-validated selection package.)

Effects of Practice on Validities. All of the correlations mentioned in the preceding paragraph were computed using data from the first hour of testing on the MTPB. As indicated previously, the rationale underlying the changes in task difficulty in Study III was based on the fact that the correlations in Study II for the second and third sessions of testing had been smaller than the correlation for the first session. We hypothesized that the decreased correlations were the result of the task complex becoming too easy with practice, and, thus, the measures lost their power to discriminate among subjects. We thought that, by increasing the difficulty of the tasks, we would increase the discriminatory power of the measures on the second or third sessions when performance would be expected to be more stable. Thus, more reliable data and higher correlations were predicted. The prediction was not accurate. Lower correlations were found on the second and third sessions in Study III and all of the correlations were near zero in Study IV. It would appear that the predictive power of the MTPB measures is best when those measures are taken during initial acquisition of skill on the task complex. This is presumed to be because the absence of the ability to shift rapidly from one task to another would have a greater impact on performance early in learning; as skill is acquired on the individual tasks, more time becomes available for attending to the other tasks. For example, the mean response times on arithmetic in session 2 were shorter than those

for session 1; hence, less time spent on arithmetic left more time for scanning the panel for meter and light signals and for working the problem-solving task. In other words, the closer the subject's "channel capacity" comes to being "used up" by the demands of the situation, the better the approximation of the measures to an index of time sharing ability. However, the task situation must not be so difficult that it is reacted to as contrived or unrealistic.

Individual Task Validities. The relative merit of the individual tasks as predictors of the criterion presents an interesting picture. The green-lights task, which was conceptually one of the simplest, showed the greatest consistency as a predictor of the instructor rating, and the red-lights task was close behind. If these tasks are considered to be just measures of reaction time, there is no convincing rationale as to why they should be particularly related to controller potential. However, if they are considered as a part of the task complex, response time to the red and green lights can be thought of as an index of time sharing ability. That is, performance on these two tasks gives us an indication as to the extent to which the individual has "spare channel capacity" when he is concentrating on the active tasks (problem solving with arithmetic or with pattern discrimination). However, these two tasks obviously do not provide a clean measure of time sharing. The subject who is fast at working arithmetic, for example, has more time available to scan the panel for light signals; this would be somewhat more important to achieving fast response on green lights than on red lights in that the attention value of the onset of a red light is greater than that of the offset of a green light.

The pattern discrimination task was the next most consistent predictor with significant correlations with the criterion in Studies II and V. Two aspects of this task may have contributed to the validity of the measures. First, the task requires the exercise of immediate memory for a visual configuration (the standard pattern), which many subjects reduced to a series of numbers corresponding to column heights. Second, and perhaps more important, the manner in which the standard and comparison patterns are presented requires the subject to be sensitive to rather critical timing intervals. The subject must generate a time base that permits him to attend to the pattern task display at the proper times, namely, during the 5-second presentation of the standard pattern and during the 2-second presentations of the comparison patterns. And of course, he must also remember which of the three patterns for a given problem he is currently viewing.

The group problem-solving task, second solution, was a good predictor in Study II, but the correlations with the instructor ratings were no significant in Studies III and IV; in Study V significant correlations were found only for the final 15 minutes when problem solving occurred with only the monitoring tasks, a condition that was not used in the earlier studies. The failure of measures from the first solution to correlate with the criterion was perhaps the result of the nature of the demands during that phase of the task. We can identify two major types of responses that occur when a new problem is presented. One type of response involves the execution of the standard search procedure in which each subject pushes his button after he sees that the preceding subject in the search sequence has pushed. As the problem is partially solved, the nature of the task changes for those individuals whose positions in the problem sequence have been found. Specifically, the correct number one man for that problem must be alert to push his button any time the red error-light is re-illuminated. The number-two man is alerted by the red light that as soon as the number-one man has pushed, he should push, etc. Thus, the nature of the task during the first solution is determined by demands that, for a given problem, are heterogeneous across subjects. Another factor is that the time at which a given subject should respond would fall more or less at random with respect to the immediate demands of the concurrently performed arithmetic or pattern discrimination tasks. Therefore, since a given response on this task was measured in terms of the elapsed time from the last preceding response by any of the subjects, an additional degree of heterogeneity is introduced. Given sufficient number of problems, this heterogeneity should balance out. (This was the major reason for increasing the rate of presentation of problems in Study V.) In contrast, the second phase places rather uniform demands on the subjects; the primary exception is that the cue for the number-one man to respond is the onset of the red light indicating

he initiation of the second-solution phase of the ask. The proportion of errors on the second olution, which showed up as a predictor only in Study II, may serve only to keep the subjects onest, i.e., it ultimately would penalize a strat-gy of responding rapidly but carelessly.

The value of the meters task remains an open uestion. We were perhaps misled by the sub-tantial, though non-significant, correlation with he criterion in Study III after the random ackground wandering was introduced into the ointers. Because the findings in Study IV were egative across the board, we did not see this as reason for returning to the stationary pointers f Study II.

The arithmetic task, per se, does not appear o contribute as a predictor, although the re-ponse-time measure approached significance in tudy II. However, the harder problems of tudies III and IV and the increased rate of resentation on the problem-solving task in tudy V preclude a firm decision on the direct alue of the task as a predictor of instructor atings of trainee potential.

In evaluating these individual validity co-ficients, one should keep in mind that the tasks a given combination are performed simul-aneously. Thus, unlike the typically-applied st selection procedures, we might very well nd it desirable or even necessary to retain tasks at do not contribute directly to the validity of e composite score. For example, as mentioned reviously, there is little in the way of convinc-g evidence that conceptually simple, choice action time tasks to the onset and offset of ghts would be of value as predictors of trainee otential. Their value presumably derives from eir being imbedded in the task complex as a hole along with, for example, the arithmetic sk.

Conversations with the instructors and with me of the advanced trainees reinforced our terpretation of the formal reactions of the in-ructors to the MTPB as seen in the evaluation rm; the expected predictive value of the TPB lies in the demands placed on the indi-dual by the task ensemble rather than in the levance of the individual tasks to the job of e controller. In other words, the ability to ne share is the critical prerequisite to becoming good controller assuming that the individual in question has demonstrated possession of ade-quate abilities on a number of specific skill di-mensions that can probably be measured satis-factorily by more traditional approaches.

Effects of Task Difficulty on Validities. Our attempts to improve the predictive power of the MTPB through various manipulations of task difficulty were probably counter-productive. Ap-parently, the task parameters that we selected for use in Study I and retained in Study II were closer to the optimum conditions than anything else we tried. The changes that were introduced in the later studies quite likely had a very un-desirable side effect. These changes seemed to push the testing situation in a direction such that it was reacted to as unrealistic as a simula-tion of the demands of the controller's job. Com-ments that suggested this possibility were overheard in Studies III and IV and, to a some-what lesser extent, in Study V. It is clear that the subjects in this kind of research must be in-duced to accept the testing situation as being believable if the results are to reflect what one might expect in an actual selection situation. (This caveat also clearly applies to *any* labora-tory research in which the attitudes of the sub-jects toward the testing situation might be expected to influence their performance.) Thus, it may well be that the subjects of both Studies III and IV reacted to the MTPB as placing unrealistic demands on them. Both groups of subjects were told that the device had shown promise in earlier studies as a predictor of suc-cess in the Academy training program. Since the subjects of Study IV were new appointees, the test situation may have been seen as very threatening; if they did not do well on the MTPB, they might very well see this as reason to expect to be eliminated from the training program. The specific effect this would be ex-pected to have on their performance is not known. On the other hand, the subjects in Study III had graduated from the Academy course at some time in the past and they knew that trainees were not eliminated from the advanced course they were attending. Hence, the MTPB test situation would likely have been less threatening to them. The hypothesized difference between the reactions of the two groups, then, would be that the performance of the Study-IV subjects was affected by the threat-producing overtones of the situation; the subjects of Study III simply

performed without regard to any implied threat. A similar process may have been at work in Study V during the second and third, high-demand test sub-periods. In this regard, the subjects of both Studies IV and V *appeared* to be much less at ease in the situation than did the subjects of Study II and Study III.

Another important difference between Studies III and IV is, of course, the fact that the subjects differed in two respects. First, as noted above, they were at different stages of their respective training programs. And, second, Study III involved En Route trainees, whereas Study IV involved Terminal trainees. The only observed performance difference between the two groups was the significantly slower responses to the meters task for the subjects of Study IV. This difference may have been a reflection of the above-mentioned hypothesis about a difference in the perceived threat of the situation. In any event, the absence of a correlation between meter performance and the instructor rating in both studies leads to the tentative acceptance of the proposition that a difference in performance on the meters task should not be considered to be diagnostic of a critical difference between the two groups as far as MTPB performance is concerned. Thus, the important difference between the two groups was the fact that a significant correlation with the instructor ratings was found for the advanced, En Route trainees of Study III but not for the entry-level Terminal trainees of Study IV. However, the conclusion that the MTPB has potential for skills relevant to the En Route controller's job but not the Terminal controller's job is not warranted, primarily because of the small sample sizes. In addition, significant correlations with the instructor ratings were found in the case of the entry-level Terminal trainees of Study V.

Although the manipulations of task difficulty provided us with some information of value with respect to the methodology of complex performance measurement, from the selection point of view, the results were a disappointment for two reasons. First, the changes did nothing to improve the predictive power of the measures—in fact they probably degraded the accuracy of the predictions. Second, the differences across studies meant that we had insufficient numbers of subjects per condition to yield satisfactory stability of the correlational analyses. The number of subjects involved in Study V was adequate for many purposes, but, as it turned out, the task configurations were apparently less than optimum.

In any future work on the MTPB as a selection device, it would appear that we should use the task characteristics of Studies I and II. Namely the meters would be used with no background noise; the easy, 2-digit arithmetic problems would be used, and the inter-solution interval or the problem-solving task would be held at 30 seconds. With respect to the task combinations we would propose to use two 30-minute test periods. The first would involve problem solving, arithmetic, and monitoring; the second would involve problem solving, pattern discrimination and monitoring.

Relation of MTPB to Written Tests. The correlations between the CSC scores and the performance scores for the sub-sets of subjects in Studies II and V suggest that the two measuring instruments sample, to a significant degree, the same kinds of behaviors. Strong support for this interpretation is found in the correlation between the CSC composite scores and the MTPB composite scores for the 145 journeyman controllers. The higher correlation between the MTPB scores and the instructor ratings for those subjects for whom we did not have CSC scores is of interest. If we assume that those subjects in Studies II and V for whom we did not have CSC scores were accepted into training without having been screened on the CSC battery, then we can infer that selection on the CSC battery tends to produce a more homogeneous group with respect to MTPB performance. Presumably, the CSC battery tends to screen out those applicants who might be expected to fall at the bottom of the distribution on MTPB performance. Thus, the prior selection, both through the CSC screening and, for studies II and III, through elimination in the training programs, resulted in an effective reduction in the range of skills exhibited by the subjects we tested. The restricted range of the samples and the small number of subjects limit the interpretive generality of a combination of the CSC and the MTPB by multiple regression analysis, but the multiple R that results with the 60 subjects of Study V is a value of .46 which is significant at better than the .01 level of co

fidence. A value this high in a group subjected to previous screening must be regarded as very substantial.

Effects of Task Difficulty on Performance. The large, consistent difference between the performance of Study-II subjects and that of the subjects of the other studies on the meter monitoring task confirms the fact that the random background motion of the pointers not only makes the task look harder, it is in fact harder. The effect of this increased difficulty on the performance of the other tasks is not easily assessed since in no case was that the only procedural difference between two groups of subjects. However, the nature of the task is such that it would have an impact on the other tasks only to the extent that extra time is devoted to scanning the meters to detect the meter signals. It is possible that the effect of the random motion would be seen only in meter performance. In this respect, it may be that the difference between the Study-II and Study-IV subjects on this task was that important numbers of the Study-IV subjects gave relatively little attention to the meters task. This speculation is confirmed by the fact that 3 of the 30 subjects of Study IV had mean response times of 24 seconds or greater whereas the mean response times of only two of the 31 subjects of Study III exceeded that value.

The increased difficulty of the arithmetic problems in Studies III and IV would be expected to have two effects. First, it would obviously affect the arithmetic task; response times would be affected since the problems now had to be solved one digit at a time; and accuracy would be affected since the step-wise solution of problems presented greater opportunity for errors. The second effect would be a product of the increased response times; the increased time devoted to arithmetic would mean that less time would be available for scanning the monitoring tasks and for attending to the problem-solving task. Both arithmetic time and accuracy were affected as seen in the comparisons of Study II with Studies III and IV. The effects on the monitoring tasks were not clear-cut; the subjects of Study III did just as well on red and green lights as did the subjects of Study II; the subjects of Study IV were slower than those of Study II on red lights but not on green lights. Both Study-III and Study-IV subjects were slower in making responses on the group problem-solving task during the first solution than were the subjects of Study II when problem solving was performed concurrently with arithmetic. In neither case were the second-solution responses slower. Interpretation of this finding on problem solving is confounded by the fact that the Study-III subjects were slower during the second solution when problem solving was performed with the pattern-discrimination task; the Study-IV subjects showed the same effect but, in addition, they were slower in responding on the first solution (problem solving plus pattern discrimination). The effects on problem solving with pattern discrimination may have been a "carry over" from a decrease in the value of the problem-solving practice during the arithmetic phase of testing.

The increased signal rate on the monitoring tasks in Study V probably induced those subjects to spend a little more time scanning the lights with a resultant improvement in performance relative to the other studies. The Study-V subjects undoubtedly developed higher expectancies with respect to the occurrence of light signals.

The decreased delay interval between solutions and between problems on the group problem-solving task in Study V produced effects not only on that task but on arithmetic and pattern discrimination as well, as seen in the comparison of Study V with Study II. The explanation is straightforward. The decrease in the inter-problem interval meant that a larger proportion of the arithmetic problems were solved while a problem was being worked on the problem-solving task. Similarly, the timing in looking at the pattern discrimination task would be more likely to be disrupted by the greater frequency of joint occurrence of pattern problems and problem-solving activity.

In retrospect, the changes in task difficulty would have to be judged as having been ill-advised. Although the effects produced were not those that were sought from the point of view of improved validity in predicting the instructor ratings, the effects were, nonetheless, substantial in nature. However, from the point of view of methodology, the difficult manipulations did provide valuable information concerning the proper task characteristics to use in further research.

V. Summary and Conclusions.

The five studies described in this report were undertaken to explore the possibility of using a performance measurement system such as the Multiple Task Performance Battery as a means for assessing the potential of trainees to become fully-rated controllers. The MTPB provides objective measures of monitoring behavior, arithmetical skills, visual discrimination, and the execution of procedures as embodied in a group problem-solving task. The important characteristic of the MTPB in relation to the job of the Air Traffic Controller is that the various tasks can be put together in combinations that impose differing work loads on the individual with respect to the demands for time sharing.

The primary criterion used in evaluating the predictive efficiency of the MTPB was the mean of three ratings for each trainee—one rating from each of three instructors per trainee. Specifically, the instructors were asked to express their opinion of each trainee's potential for becoming a fully-rated controller by assigning a number from 0 to 100 to the trainee, with 70 being "minimally acceptable potential." In addition, 40 instructors were tested on the MTPB and then asked to complete a 7-item questionnaire designed to elicit their opinions as to the relation of the MTPB to the job of the controller and the likely predictive value of the MTPB as an index of trainee potential.

The first study involved 19 En Route trainees; it served as a preliminary to the other studies in checking out training and testing procedures; it also permitted an examination of the appropriateness of the task parameters used. The second study involved 60 En Route trainees; it yielded a good validity coefficient against the criterion—product moment $r = +.54$. The third study also involved En Route trainees (a total of 31), but a different set of task parameters was employed. The validity coefficient in this study was +.53. The coefficients found in both of these studies were lower for data from the second and third hours of testing than for the first. The fourth study, which involved 30 Terminal trainees and the same task parameters as the third study, failed to replicate the validities found in the second and third studies; the correlation here was near zero. The validity coefficient found with 89 Terminal trainees in the fifth study, using still a different set of task parameters, was not as large as those of the second and third studies—$r = +.24$; however when the MTPB performance index was based on measures judged to be more appropriate in relation to the changed task parameters, the correlation was comparable—$r = +.46$.

The opinions of the instructors, as reflected in their responses on the questionnaire, support the conclusion that the MTPB exhibits substantial content validity as a measurement system of relevance to the job of the controller.

It is concluded that the task parameters used in the first study (and retained in the second study) are closer to optimum than were any of the modifications used in the third, fourth, and fifth studies. The method of computing the composite index of performance is judged to be the best technique for developing an objective measure for use as a summary assessment of performance capability on the MTPB task ensemble. This method gives equal weight to all tasks in terms of the variance of the composite index. It is also concluded that one hour of testing yields satisfactory reliability of the composite performance index.

The primary intent of these studies was to determine whether or not the MTPB approach to controller selection showed sufficient promise to warrant a full-scale evaluation of its merit. It is concluded that the proper answer to the implied question is, "Yes." The approach does appear to offer possibilities as a selection technique. However, a full-scale evaluation is required to establish the reliability of the obtained validity coefficients and to evaluate the relative contribution of the MTPB approach as a supplement to existing selection procedures in terms of cost-effectiveness.

REFERENCES

1. Brokaw, L. D.: School and Job Validation of Selection Measures for Air Traffic Control Training. USAF: WADC–TN–59–39 (AD 214–884), 1959.
2. Chiles, W. D., E. A. Alluisi, and O. S. Adams: Work Schedules and Performance During Confinement, HUMAN FACTORS, 10:143–196, 1968.
3. Cohen, Jacob: *Statistical Power Analysis for the Behavioral Sciences*, New York, Academic Press, Inc., 1969.
4. Trites, D. K.: Problems in Air Traffic Management: I. Longitudinal Prediction of Effectiveness of Air Traffic Controllers. Civil Aeromedical Research Institute Report 61–1, Oklahoma City, Oklahoma, 1961.
5. Trites, D. K., and B. B. Cobb: CARI Research on Air Traffic Control Specialists: Age, Aptitude, and Experience as Predictors of Performance. Civil Aeromedical Research Institute Report (unnumbered), Oklahoma City, Oklahoma, 1964.
6. Wallace, S. R.: Criteria for What? AMERICAN PSYCHOLOGIST, 20:411–417, 1965.

APPENDIX I—Instructor's Evaluation Questionnaire of MTPB.

1. To what extent do you feel the work load demands of the Multiple Task Performance Battery to be similar to those imposed on an En Route (Terminal) controller during some portion of his normal duties?

Not at all like						Very much like

2. What percent of the time does he experience that work load?_____

3. Is performance on the Multiple Task Performance Battery more or less difficult than the job of a controller?

Very much less difficult						Very much more difficult

4. To what extent do you think that the behavioral functions measured by the Multiple Task Performance Battery are important to proficiency as a controller?

Very unimportant						Very important

5. Rank the tasks from best (1) to worst (6) as predictors of the potential of a trainee to become a good controller.

_____ red lights
_____ green lights
_____ meters
_____ arithmetic
_____ problem solving
_____ pattern discrimination

6. Rank the task combinations from best (1) to worst (4) as predictors of the potential of a trainee to become a good controller.

_____ arithmetic, meters, lights
_____ arithmetic, problem solving, meters, lights
_____ patterns, problem solving, meters, lights
_____ patterns, meters, lights

7. If you had to make a decision about a marginal trainee, do you think that you would find information on the performance of these tasks helpful?

of very little use						of very great use

APPENDIX II—Instructor's Evaluation of Subjects (Ratings).

INSTRUCTOR RATING SHEET

Please circle the number below that indicates *your* opinion of the potential of _____ for becoming a
(name of student)
fully rated air traffic controller. For purposes of this evaluation each instructor should give his *own* opinion without knowledge of other instructor ratings.

Insufficient potential	0	30	40	45	50	55	56	57	58	59	
Questionable potential	60	61	62	63	64	65	66	67	68	69	
Below Average	70	71	72	73	74	75	76	77	78	79	
Average	80	81	82	83	84	85	86	87	88	89	
Above average	90	91	92	93	94	95	96	97	98	99	100

APPENDIX III—Correlations Between Performance Composite and Written Tests*.

CSC	
Arithmetic Reasoning	.29**
Spatial	.22**
Following Oral Directions	.34**
Abstract Reasoning	.35**
Air Traffic Problems	.38**
Composite	.44**
COINS Test	.32**
Dailey Technical and Scholastic Test	
Electricity	.12
Electronics	.17*
Mechanical	.20**
Physical Science	.19**
Arithmetic Reasoning	.31**
Algebra	.16*
Vocabulary	.30**
General Information	.11
Air Force Dial and Table Reading	
Dial Reading	.30**
Table Reading Composite	.42**
CAMI Directional Headings Composite	.38**

*p<.05
**p<.01

* Data obtained on 183 ATCSs under Contract DOT-FA70WA-2371.

21819