Summer 1986

# The Development of a Paper-and-Pencil Measure of Complex Cognitive-Perceptual Aptitude

Don Michael McAnulty
*Old Dominion University*

# THE DEVELOPMENT OF A PAPER-AND-PENCIL MEASURE OF COMPLEX COGNITIVE-PERCEPTUAL APTITUDE

by

Don Michael McAnulty
B.S. August 1978, Old Dominion University
M.S. December 1980, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

PSYCHOLOGY

OLD DOMINION UNIVERSITY
July, 1986

Approved by:

Glynn D. Coates (Director)

## ABSTRACT

THE DEVELOPMENT OF A PAPER-AND-PENCIL
MEASURE OF COMPLEX COGNITIVE-PERCEPTUAL APTITUDE

Don Michael McAnulty
Old Dominion University, 1986
Director: Dr. Glynn D. Coates

The primary purpose of the present research was to develop a complex aptitude test to assess individual differences in multiple cognitive and perceptual abilities that are required for helicopter pilot training. The paper-and-pencil test was designed to provide measures of both static and dynamic (i.e., learning) ability under different levels of complexity. The secondary purpose of the research was to develop a battery of eight psychometric tests to assess other abilities that are required for helicopter pilot training. A prototype of each test was produced and administered to small samples of subjects during the preliminary phase of testing. The tests were then revised and compiled into an experimental battery that required approximately 7 hours to administer. The battery was administered on six test dates to 290 subjects at three military bases. Seventeen of the subjects were subsequently deleted from the analyses for failing to provide complete data, failing to follow test procedures, etc. Performance on the experimental battery by the remaining 273 subjects is considered to be representative of the broad spectrum of abilites that exists among military personnel who are in the training stage of their careers. The psychometric characteristics of most of the tests are sufficient to justify

further research on their utility as selection instruments. The average difficulty levels are near the optimum level of .50, the test variances indicate the measurement of substantial individual differences, and the estimates of reliability are acceptable when test length and the design specifications are considered. Factor analyses indicate that the battery assesses seven independent dimensions of human abilities. When the test sections are combined into total scores, the battery assesses three primary ability domains: cognitive-perceptual, perceptual speed, and spatial orientation. Corrections for guessing did not substantially alter the underlying factor structure of the battery. Two additional factors were extracted when the within-difficulty-level gain scores from the complex test were included in the total score analysis. The gain scores at the lower level of difficulty are interpreted to be indices of learning while the gain scores at the higher level of difficulty are interpreted to be indices of fatigue. It was concluded that the complex test and six of the psychometric tests assess reliable individual differences in the abilities of interest. A validation study was recommended to determine if these tests, with minor modifications, are predictive of student performance in helicopter flight training.

## DEDICATION

Many people have influenced my development as a person and as a psychologist. Several will be acknowledged on the following page for their influence and their contributions to this project. However, this dissertation and the effort that it represents is dedicated to the most important people in my life:

First of all, it is dedicated to the memory of my parents, Mac and Tiny. Although they are not here to share in this achievement, my parents gave me life, nurturance, guidance, opportunity, and a desire to learn. Without them, neither I nor this work would exist.

Second of all, it is dedicated to my children, Becky, Timmy, and Jesse, who have opened a new dimension of life to me. I hope to give them the nurturance, guidance, opportunity, and desire to learn that I received. Mostly, I hold the vain hope that my accomplishments and those of my wife will someday serve to elevate their goals and stimulate their development.

Most of all, it is dedicated to my wife, Beverly. It is impossible to enumerate all the ways that Bev has influenced me and contributed to this project. Neither I nor it would be the same without her unflagging support, stimulation, and confidence that this goal would eventually be accomplished. She certainly shares the credit, and I trust will share the benefits of its completion.

## ACKNOWLEDGMENTS

I would like to acknowledge and thank the members of my Ph.D. guidance and dissertation committees, Dr. Glynn Coates, Dr. Raymond Kirby, and Dr. Elaine Justice. Their mentorship throughout my educational process has been a source not only of guidance and tutelage, but also a source of inspiration and a model of professional excellence. Very special appreciation is extended to Dr. Coates and Dr. Kirby who also guided my Master's degree studies and supervised my research and teaching activities.

I would also like to thank Dr. Kenneth Cross and Anacapa Sciences, Inc. Dr. Cross, a member of my dissertation committee and Vice President of Anacapa Sciences, has fully supported and actively, yet patiently, encouraged my efforts to complete this degree. Furthermore, Dr. Cross and my colleagues at Anacapa have provided models of professionalism in an applied environment that is sometimes more conducive to quantity than to quality.

In particular, I would like to acknowledge Dr. Dennis Jones, my colleague at both Old Dominion and Anacapa Sciences, who was irrepressible in his encouragement (some would call it nagging) to complete this research.

Once again, I would like to acknowledge the contributions of my wife Beverly. She has provided intellectual stimulation, constructive criticism, unwavering confidence in my ability, consistent encouragement (I wouldn't dare call it nagging), and the time and space required to complete this work.

Finally, I would like to acknowledge the sponsorship of the U. S. Army Research Institute Aviation Research and Development Activity, Fort Rucker, AL. This research could not have been accomplished without their support. However, the views, opinions, and findings described in this report are those of the author and should not be construed as an official Department of the Army position, policy, or decision.

TABLE OF CONTENTS

## LIST OF TABLES

## Introduction

The use of cognitive and perceptual aptitude tests has significantly improved personnel decision-making in many organizations (Schmidt & Hunter, 1981). Recent evidence has shown that the use of valid selection procedures has increased worker productivity and improved the economic performance of the employing organization (e.g., Schmidt & Hunter, 1983; Schmidt, Hunter, McKenzie, & Muldrow, 1979; Schmidt, Hunter, & Pearlman, 1982). Further studies have shown that the general utility of selection tests across various circumstances and populations is more extensive than had been presumed. For example, a number of validity generalization studies (e.g., Schmidt, Hunter, Pearlman, & Shane, 1979) on selection testing have mitigated the long-held presumption of situational specificity (e.g., Freyd, 1923; Ghiselli, 1966), although these meta-analytic conclusions are not universally accepted (e.g., Schmitt, Gooding, Noe, & Kirsch, 1984). Finally, empirical evidence collected over the last two decades has not supported earlier concerns about the single-group or differential validity of selection tests (e.g., Hunter & Hunter, 1984).

Despite the positive results of these studies, the overall success of aptitude testing for employment, training, and educational selection has not been overwhelming. Although the test validities have been high enough to justify the purported increases in worker productivity, the average criterion-related validity of selection tests has been relatively low (Boehm, 1982; Ghiselli, 1973; Schmitt, et al., 1984). A number of factors have been shown to affect the validity coefficients of selection tests. Cumulatively, the validity generalization studies (e.g., Schmidt, Pearlman, Hunter, & Hirsh, 1985) indicate that many findings of low validity are attributable to statistical and procedural artifacts such as restriction in range, inadequate sample sizes, criterion unreliability, and administrative errors that limit the maximum correlation that can be obtained. In their review, Schmitt, et al., (1984) found differences in the average validity coefficient as a function of the research design (e.g., restriction in range when the predictor test is used for selection in the validation study), the occupational group (e.g., sales and skilled labor groups typically have the lowest validity coefficients), the predictor type (e.g., personality measures have the lowest validity estimates), the criterion type (e.g., performance ratings yield higher validity coefficients than turnover or productivity criteria), and the predictor-criterion relationship (e.g., educational grades are best predicted by mental ability tests and most poorly predicted by personality tests).

The correspondence between the predictor and criterion is particularly important in the development of valid tests. For example, Gutenberg, Arvey, Osburn, & Jeanneret (1983) found that the predictive validity of cognitive tests was positively moderated by the degree that decision making and information processing were required. The cognitive test scores were significantly correlated with performance on the jobs that had high levels of cognitive requirements but they were not correlated with performance on the jobs that had low levels of cognitive requirements. These results are consistent with the behavioral consistency approach expounded by Wernimont & Campbell (1968).

In a meta-analytic review of validity studies, Hunter and Hunter (1984) found that work sample tests were best for selection based on current job performance (i.e., selecting among experienced applicants) but that cognitive aptitude tests were nearly as predictive for entry-level jobs. The statistical validity of cognitive aptitude tests generally improved as the complexity of the training or job increased. Hunter and Hunter concluded (p. 80) that "There is no job for which cognitive ability does not predict training success." Although the predictor-job performance relationship is moderated by the requirements of the job, the training (i.e., learning) requirements for entry-level positions are best predicted by cognitive aptitude tests, regardless of the subsequent performance requirements. Narrative reviews (e.g., Guion, 1976) of validity studies that compare training and job-performance criteria also support this inference.

Approaches to Ability Testing

Although cognitive and perceptual ability tests have been highly successful in predicting training and job performance, there is a need for improvement in the proportion of criterion variance accounted for by the predictor. The contemporary emphasis in ability testing is on the assessment of information processing capabilities (e.g., Embretson, 1985; Sternberg, 1985). This approach attempts to analyze the mental representations and component processes that underlie cognitive operations. There are numerous research paradigms (e.g., Carroll, 1976, 1978; Hunt, 1980; Sternberg, 1983) within the information processing approach that vary in their levels of analysis and in the complexity of their theories. Several researchers (e.g., Barrett, Alexander, Cellar, Doverspike, & Thomas, 1983; Imhoff & Levine, 1981) have attempted to develop information processing test batteries for applied purposes, but the batteries have not yet been implemented extensively. Although the information processing approach toward the measurement of individual differences is promising, it is still in the experimental stages of development (e.g., Carroll, 1978; White, 1985). The approach has not yet developed methods for substantially improving the prediction of training or job performance (e.g., Bejar, 1985).

The traditional and predominant approach to applied aptitude testing has been to develop a battery of psychometric tests that measure multiple, individual abilities. Psychometric tests are designed to assess individual differences in the structure rather than in the processing of human abilities. Ability measures are then combined statistically to predict complex training or job performance (e.g., Cronbach, 1970; Guion, 1965, 1976). For example, one of the most widely used test batteries is the Armed Services Vocational Aptitude Battery (ASVAB) which is administered to all enlistees in the armed services. The ASVAB is composed of ten, individually administered subtests (e.g., Arithmetic Reasoning, Word Knowledge, and Coding Speed). Scores from the subtests are combined in various ways to produce composite scores that are used for initial selection (the Armed Forces Qualification Test score) or assignment to training slots or career fields (clerical, maintenance, communications, etc.). The current area composites have been validated against training and skill qualification test scores for a large number of military occupational specialities (McLaughlin, Rossmeissl, Wise, Brandt, & Wang, 1984).

Critique of the Psychometric Approach

There are several unresolved issues within the psychometric approach to applied selection testing. The primary issues include the theoretical assumptions of the approach, the types of ability constructs that are measured, the administrative format of the tests, and the procedures used to integrate information from a battery of tests.

Two critical assumptions underlying the psychometric approach have been criticized as being untenable (e.g., Allen, Secunda, Salas, & Morgan, 1982). First, the psychometric approach assumes that all examinees have had an equal opportunity to benefit from life experiences prior to testing. In a heterogeneous sample such as all military accessions, this assumption is highly questionable. Considering the diversity of backgrounds of the large number of military recruits, the opposite assumption is, in fact, more likely to be valid. Second, the psychometric approach assumes that the test scores, which are static measures of current ability, are predictive of the capacity to acquire and use information and skills, which are dynamic processes. The second assumption is dependent upon the first. That is, if the examinees have not had an equal opportunity to acquire knowledge and skill, then static measures of achievement must be deficient predictors of cognitive aptitude. This conclusion is consistent with theoretical (e.g., Wernimont & Campbell, 1968) and empirical (e.g., Hunter & Hunter, 1984) evaluations of the predictor-criterion relationship.

As an alternative to the static measurement of abilities, Christal (1976) has suggested that learning rate measures on cognitive tasks may be useful predictors of training and job performance. The learning rates would provide for the dynamic measurement of abilities because they would assess changes in performance rather than levels of performance. Gettinger & White (1979), for example, found that the number of trials to criterion on a learning task was a better predictor of school achievement than a group-administered intelligence test. Several research projects (e.g., Allen & Morgan, 1984; Payne & Tirre, 1984) have recently examined the possibility of measuring learning aptitude directly.

These projects are usually designed for computer-based tests rather than for paper-and-pencil tests (e.g., Christal, Tirre, & Kyllonen, 1984). Compared to paper-and-pencil tests, computer-based tests can assess a broader range of abilities (e.g., psychomotor abilities) and can use a broader range of measures (e.g., response time). In addition, computerized tests can be adapted to each subject's ability level, thus reducing the time required to administer the tests (e.g., Bejar, 1985). Despite these advantages, computer-based tests are not widely used in applied settings. They are more expensive to develop or acquire, and more difficult to maintain. The logistics of computerized testing become more problematic when a large number of individuals must be tested at a large number of locations. While instrumentation may be preferred for psychomotor tests (e.g., Guion, 1965) or work sample tests (e.g., Cascio & Phillips, 1979), the majority of large-sample studies have shown that paper-and-pencil tests are excellent measures of ability and that other types of tests are usually more expensive and less valid (Hunter & Hunter, 1984). Furthermore, Smith, Krause, Kennedy, Bittner, & Harbeson (1983) found that computerized tests may not be as reliable as paper-and-pencil tests and that the test formats may measure different behavioral constructs. It is for these reasons that most organizations continue to use paper-and-pencil instruments as their primary source of aptitude assessment in applied settings. It is apparent that concurrent efforts should be directed toward the measurement of learning ability that is relatively free of the effects of past experience using a practical test format.

Finally, the psychometric approach of statistically combining measures of multiple abilities may not be an optimal strategy for predicting complex behavior. The primary problem in combining measures of unique abilities into a composite score is the determination of appropriate weights for each measure. One method is to assign equal weights to each measure. Unit weighting is appropriate when the ability requirements of a complex task are equivalent. More often, however, it is chosen by default when there is no evidence of a clear inequality of ability requirements or when there is no reliable procedure to determine the relative ability requirements.

The most common alternative to unit weighting is the use of multiple regression analyses to determine appropriate weights for each measure. The critical problem with this alternative is that the regression weights are derived from a sample but applied to individuals. Complex behavior is unlikely to be predicted for all individuals in the same way (Guion, 1976). That is, individuals will utilize their stronger abilities to compensate for their weaker abilities when performing a complex task. Hunt (1976) has argued that the performance of any two individuals on the same task depends upon different, non-linear combinations of their structural capacities. For example, the cognitive-processing strategy that an individual chooses may be influenced by both the capabilities of the individual and the complexity of the training or job task. Onken, Hastie, & Revelle (1985) found individual differences in the processing of information on a cognitive task as the level of complexity increased. Some subjects exhibited a linear increase in processing time as the task became more complex while other subjects exhibited a curvilinear relationship between processing time and complexity. The latter function indicates that these subjects either adopted a strategy to simplify the task or altered the parameters or criteria for making decisions at the highest levels or complexity.

In summary, the measurement of complex behavior that requires the use of multiple abilities may be more predictive of future complex performance than the measurement and statistical combination of multiple, simple abilities (e.g., Asher & Sciarrino, 1974; Wernimont & Campbell, 1968). Furthermore, the assessment of learning ability may diminish the effects of individual differences in experience and improve the correspondence between the predictor and the training criteria.

Purposes of the Study

The primary purpose of the present research was to develop a complex cognitive-perceptual ability test that fulfills the following requirements:

- assesses individual differences in multiple cognitive and perceptual abilities,
- provides measures of both static and dynamic (i.e., learning) ability,
- measures aptitude under different levels of complexity,
- minimizes the effects of prior learning and experience,

- maintains a reasonable degree of "face" validity to applicants in an applied setting, and
- meets the logistical requirements (e.g., reasonable time limits, group administration, paper-and-pencil format) for large scale use.

The complex test was developed as a potential predictor of performance in the U. S. Army Initial Entry Rotary Wing (helicopter) flight training program. Flying has long been recognized as a complex training and performance environment. In 1966, Passey and McClaurin described three categories of aircrew characteristics that are important for successful flight operations: (a) adapatability to changing situations, (b) capacity for integration and processing of information from multiple sources, and (c) storage, reorganization, comparison, and combination of data input. Technological advances in the last two decades have decreased the psychomotor workload required to control the aircraft but have increased the pilot's decision-making workload (Jensen, 1982). Imhoff and Levine's (1981) review of the literature identified an extensive list of perceptual, cognitive, and psychomotor abilities that are required for effective pilot behavior. Based on that review, a computerized battery of psychomotor, cognitive, and personality tests is currently being developed to predict pilot training performance in the U. S. Air Force (Kantor & Bordelon, 1985).

The majority of studies and reviews have focused on fixed-wing pilot training and performance. Recent research (McAnulty, Jones, Cohen, & Lockwood, 1984) identified 24 abilities that are required for successful rotary-wing pilot training. Eleven of the ability constructs were chosen for selection test development on their amenability to a multiple choice, paper-and-pencil format and on their potential for reliable, nonredundant measurement. Therefore, a secondary purpose of this research was to develop a battery of tests, including the complex test, that was designed to measure the following abilities: decision making, deductive reasoning, inductive reasoning, information ordering, flexibility of closure, memorization, perceptual speed, problem sensitivity, selective attention, spatial orientation, and speed of closure.

In addition to providing a traditional psychometric framework for interpreting the complex test, the development of the battery fulfilled the operational requirements of the supporting organization. Technological advances in aircraft, increases in the complexity of missions, changes in the applicant population, and developments in test methodology necessitate the revision of the current selection battery. The current battery is a modified version of paper-and-pencil tests that were developed in the 1955-1965 time frame. The paper-and-pencil format continues to be a pragmatic requirement since several thousand applicants are tested annually at U. S. Army installations worldwide.

The complex test is imbedded in a flight planning context that maintains the appearance of relevance to aviation applicants. The test employs common concepts and terms (e.g., distance in miles, speed in miles per hour) that most, if not all, applicants should have experienced. The test obtains repeated measures at the same and at different levels of cognitive-perceptual workload. This design permits the measurement of both static ability and dynamic learning (change in performance as a function of study and practice) within the same level of information load, and the assessment of the decrement, if any, that results from imposing a higher level of cognitive-perceptual demand.

The evaluation of the complex test and the psychometric battery addressed the following questions:

(1) What are the demographic characteristics of the current sample of subjects? Are the backgrounds of the subjects as heterogenous as

would be expected from the large population of military personnel? What inferences can be drawn from this sample of subjects about potential applicants for helicopter flight training?

(2) Are the psychometric characteristics of the tests satisfactory in terms of the generally accepted standards for difficulty, variability, reliability, and discriminability? Do the psychometric characteristics of the tests warrant further research in a validation study?

(3) Are there learning effects between sections of the tests in the battery? In particular, does learning occur as a function of study and practice within the levels of difficulty on the complex cognitive-perceptual test?

(4) What are the psychometric characteristics of the learning and complexity indices that were derived from the complex test? Can, as Christal (1976) has advocated, learning be reliably assessed as a dynamic ability measure for predicting future learning performance? What effect does the level of difficulty have on the learning measures?

(5) What is the factor structure of the battery of aviation-related tests? Do the tests assess independent abilities? Does the factor structure change as a function of the various measures that were derived from the tests?

(6) Is the complex test associated with multiple cognitive and perceptual abilities? Are the dynamic measures of performance independent of the static measures of performance?

## Method

The development of the complex, cognitive-perceptual test and the battery of psychometric tests was conducted in two phases. During the preliminary test phase, groups of tests were developed and administered to small samples of subjects. The results of each administration were used to revise these tests while additional tests were being developed. The preliminary test phase continued until all the tests in the battery had been administered and revised at least three times. During the experimental test phase, the battery was administered to samples drawn from a heterogenous population of military personnel. Descriptions of the tests that were included in the experimental battery are presented first to facilitate the description of the preliminary and experimental test procedures.

### Test Descriptions

The tests that were included in the experimental battery are listed in Table 1 and described in more detail below. Copies of the instructions for the nine developmental tests are presented in Appendixes A through I. Definitions of the ability constructs were excerpted from the Ability Requirements Scales (e.g., Fleishman, 1975; Fleishman & Quaintance, 1984) that were used to analyze the requirements for initial entry helicopter training (McAnulty, Jones, Cohen, & Lockwood, 1984).

Flight Planning Test (FPT). The FPT was designed as a complex, cognitive-perceptual test to assess decision making, memorization, and selective attention abilities in an aviation related context. These abilities were defined as ". . .the ability to select the most effective course of action after considering different options and potential outcomes", ". . . the ability to remember information, such as words, numbers, pictures, and procedures . . . by themselves or with other pieces of information", and ". . . the ability to

Table 1

Descriptive Information on the Tests in the Experimental Battery

| Test Name | Acronym | Sections | Items[a] | Time[a] | Series[b] |
|---|---|---|---|---|---|
| Flight Planning Test | FPT | 7 | 12 | 9 min[c] | 9 |
| Chart Use Test | CUT | 5 | 6 | 4 min[d] | 1 |
| Map Planning Test | MPT | 3 | 20 | 3 min | 6 |
| Sound Reasoning Test | SRT | 2 | 16 | 6 min | 2 |
| Finding Rules Test | FRT | 2 | 16 | 8 min | 2 |
| Rapid Match Test | RMT | 2 | 48 | 4 min | 2 |
| Figure Orientation Test | FOT | 2 | 42 | 5 min | 2 |
| Finding Figures Test | FFT | 2 | 60 | 5 min | 2 |
| Obscured Figures Test | OFT | 2 | 28 | 2 min | 2 |
| | | | | | |
| Identical Pictures Test | IPT | 2 | 48 | 1.5 min | 1 |
| Card Rotations Test | CRT | 2 | 80 | 3 min | 1 |
| Hidden Patterns Test | HPT | 2 | 200 | 3 min | 1 |
| Gestalt Completion Test | GCT | 2 | 10 | 2 min | 1 |

[a] per section.

[b] number of order-of-presentation series; 1 = a fixed order of presentation.

[c] a 30-second study period is also provided before each section.

[d] a 2-minute study period is also provided before each section.

concentrate on a task one is doing and not be distracted", respectively. To a lesser extent, the FPT was also designed to require the other IERW abilities for maximum performance. Because of the complex nature of the FPT, highly detailed test specifications (see Appendix J) were prepared to control the equivalence of the sections within difficulty levels and to specify the differences between levels.

The FPT is a series of seven, separately timed sections divided into three levels of difficulty (see Appendix A). Each section contains a route map and 12 questions about flights between two points on the map. Instructions are presented at the beginning of each difficulty level. The basic information that is required for all seven route map sections is presented before the first difficulty level. The basic information includes: (a) the general form of the route maps, (b) the coordinate system, (c) the compass directions, (d) the airfield and landmark symbology, (e) the approved air route system, and (f) the fundamental rules of route selection. In order of importance, the fundamental rules are: (a) select the shortest route, (b) select the route that requires the fewest turns, and (c) select the route that passes the most landmarks.

The instructions also provide three practice questions followed by feedback on the correct responses and reasons for selecting them. Following the practice questions and a 30 second study period, the single, basic difficulty level, test section is administered. Three types of questions (how many turns are required . . ., what compass headings are required . . ., and which landmarks would you pass in flying from ___ to ___?) are presented at the first difficulty level.

The second level of difficulty presents instructions and practice questions on two additional variables, distance and airspeed, that are used on the remaining six sections. The horizontal and vertical distance between any two coordinate points on each map is defined as 10 miles. The airspeed variable may be presented numerically (e.g., 100 mph) or symbolically using an airspeed indicator. The mph values on the airspeed indicator must be memorized since they are not labeled on the test questions.

Two formulas are presented for determining the required airspeed and the required time to fly between any pair of locations. The formulas must also be memorized. Following the instructions and practice questions, three route map test sections are presented that require processing of the basic and second level information. A 30-second study period is permitted between each route map to review the instructions. Three additional types of questions (how many minutes are required . . ., what airspeed is required . . ., and how many miles are flown from ___ to ___?) are presented at the second difficulty level. On one third of the questions, irrelevant information on airspeed or time is included to provide selective attention distractions.

The third level of difficulty adds flight altitude information and flight restrictions to the approved air routes. The flight restrictions are presented symbolically on the route map. Flight altitude information may be presented verbally (e.g., at high altitude) or symbolically as an altimeter. The meanings of the two flight restriction symbols and the altitude values on the altimeter must be memorized. The two variables interact in that the flight restrictions are relevant only at certain altitudes. Following the instructions, three route map sections are presented that require processing of the basic, second and third levels of information. A 30-second study period is permitted between each route map section to review the instructions. The six question types used at the second level of difficulty are also used at the third level of difficulty, with the addition of altitude information being presented as part of each question.

The difficulty of the items in each section within a level of difficulty was balanced in terms of route length, number of turns, number of landmarks, route terminals, flight directions, response terms, and distracting information (see Appendix J). The order in which the items were presented was randomly determined with the constraints that the end point of one item could not be the start point of the next item and that a question type could not be repeated until each of the other question types had been presented. The difficulty levels of the FPT were presented in a fixed order. Within each of the second and third difficulty levels, the order of presentation of the three route map sections was completely counterbalanced. The presentation series in difficulty levels 2 and 3 were also fully crossed, resulting in nine orders of presentation (between 27 and 33 subjects completed each order).

Chart Use Test (CUT). The CUT was designed to measure information ordering, defined as ". . .the ability to follow correctly a rule or set of rules to arrange things or actions in a certain order. The rule or set of rules to be used must already be given." Each section of the CUT also requires other perceptual and cognitive abilities for maximum performance, but the test sections are not cumulative. The CUT is divided into five, individually timed sections that are based on charts described in the utility helicopter operators manual. Each section begins with a practice chart and two demonstration questions (see Appendix B). Two minutes are allowed to study the practice chart and the instructions for answering the questions. Following the practice period, the same chart and six test questions are presented for a 4-minute test period.

During the experimental administration of the CUT, the five sections were presented in a fixed order of increasing difficulty. On each chart section, the six test questions were arranged in order of ascending difficulty. The sections, in order of presentation, are:

- "Temperature Conversion" (a Celsius to Fahrenheit conversion chart),
- "Climb-Descent" (a chart depicting the relationship between airspeed and the rate and angle of climb or descent),
- "Fuel Load" (a chart depicting the relationships between the volume and weight of two grades of fuel and an arbitrary variable representing the aircraft balance),
- "Operating Limits" (a chart depicting the relationships between altitude, temperature, aircraft weight, and airspeed), and
- "Hover" (a chart depicting the relationships between altitude above sea level, temperature, aircraft weight, altitude above ground level, and an arbitrary variable representing the aircraft power level (i.e., torque).

Map Planning Test (MPT). The MPT was designed to assess problem sensitivity, defined as ". . . the ability to tell when something is wrong or is likely to go wrong." Test specifications were originally drafted under contract by Hills, Douglas, and Lassiter (1983), but they were substantially modified during the development of the test. The MPT is divided into three, separately timed sections containing 20 questions each (see Appendix C). Each section contains two maps, with ten test items associated with each map. Each map is composed of eight vertical and eight horizontal lines representing "streets." Circles that are placed on some of the streets indicate blocked routes. There are numbered boxes, representing buildings, placed at ten street intersections on the interior of each map. The intersection of the vertical and horizontal lines on the perimeter of the maps are identified by the letters A through Z. The task on each item is to determine the shortest route between two points on the perimeter and to identify the selected route by the building number that the route passes. The best route between the two points on any item passes one and only one building.

The MPT is, at least superficially, a less difficult version of the FPT (e.g., test questions are based on the selection of a preferred route on an abstract map). Although the MPT probably requires some additional perceptual and cognitive abilities, the memory and selective attention components of the FPT are not incorporated in the test. In addition, there are fewer decision rules to consider, fewer information variables to integrate, and no computational requirements.

On the other hand, the MPT is not an exact operational definition of the problem sensitivity construct. The problem stems from the lack of a well developed construct. The same marker tests that have been used to define problem sensitivity have also been used to define the ability constructs of perceptual foresight and spatial scanning (Hills, et al., 1983). In addition, problem sensitivity is not readily reconciled with the structure of the intellect model (Guilford & Hoepfner, 1971). However, Hills, et al. reviewed approximately 25 test models and selected the MPT model as the most appropriate measure of the ability required in IERW.

The three sections of the MPT were designed to be equivalent. The difficulty of the items was controlled in terms of the length of the routes and the number of obstacles that were encountered along the routes. The items were randomly distributed in each section, with the restriction that the end point of one route could not be the start point of the next route. For experimental purposes, the order of presentation of the sections was counterbalanced, resulting in six presentation series (between 44 and 51 subjects completed each series).

Sound Reasoning Test (SRT). The SRT was designed to assess deductive reasoning, defined as ". . . the ability to apply general rules to specific problems to come up with logical answers." Test specifications for the SRT were originally prepared under contract by Hills, et al. (1983), but only the general approach of using syllogisms to test for deductive reasoning was retained. Syllogistic reasoning is widely used as a measure of deductive ability (e.g., Sternberg, 1983).

The SRT is a multiple choice test of syllogistic reasoning that uses nonsense (consonant-vowel-consonant) syllables as the major and minor terms (see Appendix D). The nonsense syllables were selected to be pronounceable but meaningless. Each consonant was used approximately an equal number of times and each vowel was used approximately an equal number of times. The test is divided into two, individually timed sections of 16 questions each. Each question presents a major and minor premise followed by a choice of four conclusions. The level of difficulty in each section is equated by selecting items that are equivalent in terms of the figure and mood of the syllogism. The order of presentation of the sections was counterbalanced across subjects ($n_1$ = 136; $n_2$ = 137).

Finding Rules Test (FRT). The FRT was designed to assess inductive reasoning, defined as ". . . the ability to combine separate pieces of information, or specific answers to problems, to form general rules or conclusions. This includes the ability to think of possible reasons why things go together." The original test specifications were prepared under contract by Hills, et al. (1983) but they were substantially modified in the initial version of the test. The FRT also required several revisions during the preliminary phase of testing.

The FRT is divided into two, individually timed sections of 16 questions each (see Appendix E). Each question on the FRT presents four independent lines containing 20 elements. Each element may be a dot, a circle, or a space that divides elements into groups. The single circle has been placed on each line in accordance with a rule or set of rules (examples of the types of rules and the maximum number of rules are presented in the instructions). The subject must study the four lines on each question to determine the rule or set of rules that have been used to place the circle. The subject must then indicate the appropriate position for the placement of a circle on the fifth (test) line. The level of difficulty in each section is equated by controlling the number of rules that are required for each question. The order of presentation of the sections was counterbalanced across subjects ($\underline{n}_1 = 138$; $\underline{n}_2 = 135$).

Rapid Match Test (RMT). The RMT was designed to assess perceptual speed, defined as ". . . the degree to which one can compare letters, numbers, objects, pictures, or patterns, both quickly and accurately." The RMT was modeled after the Identical Pictures Test (IPT) in the Educational Testing Service (ETS) kit (Ekstrom, French, & Harman, 1976). The IPT was administered as part of the experimental battery.

The RMT is divided into two, individually timed sections containing 48 questions each (see Appendix F). Each question presents a target figure followed by four alternative figures. In the "Identical " section, the subject must determine which of the four alternative figures is identical to the target figure. In the "Different" section, the subject must determine which of the four alternative figures is not identical to the target figure (cf. Farell, 1985). The same target figures are used in both sections. The order of presentation of the sections was counterbalanced across subjects ($\underline{n}_1 = 138$; $\underline{n}_2 = 135$).

The primary difference between the RMT and the IPT is the different instructions for the two sections (both sections of the IPT require the subject to determine which figure is identical to the target). In addition, the RMT and IPT differ in terms of the number of alternatives (four versus five), the number of questions (96 versus 48), the time allowed (4 versus 1.5 minutes per section), the answer format (marking a separate answer sheet versus marking on the test sheet), and the complexity of the figures employed in the tests (a range from fairly simple to highly complex on the RMT versus all figures being fairly simple on the IPT).

Figure Orientation Test (FOT). The FOT was designed to assess spatial orientation, defined as ". . . the ability to tell where you are in relation to the location of some object or to tell where the object is in relation to you." The FOT was modeled after the Card Rotations Test (CRT) in the ETS kit. The CRT was also administered as part of the experimental battery.

The FOT is divided into two, individually timed sections containing 42 questions each (see Appendix G). On each question, a target figure is presented and followed by four alternative figures which are rotated in a random order by 45, 135, 225, or 315 degrees from the target figure axis. In the "Rotated" section, the subject must determine which of the alternative figures is only a rotated version of the target figure; the other three alternative figures are rotated and inverted (i.e., flipped over). In the "Inverted" section, the subject must determine which of the alternative figures is both a rotated and inverted version of the target figure; the other three alternative figures are only rotated. Except for variations in their initial orientation, the target figures are identical in the two sections and are presented in the same order. The order of presentation of the sections was counterbalanced across subjects ($n_1$ = 135; $n_2$ = 138).

The primary differences between the FOT and the CRT are the different instructions for the two sections of the FOT (both sections of the CRT require the subject to indicate "same" if the figure is only rotated and "different" if the figure is both rotated and inverted) and the stimulus format (in the CRT, a target figure is followed by eight alternative figures; the subject must make a judgment of "same" or "different" for each figure). The tests also differ in the number of items (84 versus 80), the time allowed (5 versus 3 minutes per section), the answer format (marking a separate answer sheet versus marking on the test sheet), and the complexity of the figures used in the tests (a range of fairly simple to highly complex on the FOT versus all figures being fairly simple on the CRT).

Finding Figures Test (FFT). The FFT was designed to assess flexibility of closure, defined as ". . . the ability to identify or detect a known pattern (like a figure, word, object) which is hidden in other material. The task is to pick out the pattern you are looking for from the background material." The FFT was modeled after the "Hidden Patterns Test" (HPT) in the ETS kit. The HPT was administered as part of the experimental battery.

The FFT is divided into two, individually timed sections containing 60 questions each (see Appendix H). On each question, the subject must determine which of four alternative patterns contains the target figure. In one section, the target figure resembles the outline of an arrowhead; in the other section, the target figure resembles an inverted question mark. The latter section was designed to be more difficult since the target figure could be located on the left, center, or right of the pattern. The arrowhead target figure could only be located in the center of the pattern. The target figures within the patterns are always in their original orientation. The order of presentation of the sections was counterbalanced across subjects ($n_1 = 137$; $n_2 = 136$).

The primary differences between the FFT and the HPT are the design of the target figures (the HPT target figure resembles an upside-down Y), the number of target figures (two versus one), and the response format (determining which of four patterns contains the target figure versus judging whether each pattern does or does not contain the target figure). In addition, the tests differ in the number of questions (120 versus 400), the time allowed for each section (5 versus 3 minutes), and the answer format (marking a separate answer sheet versus marking on the test sheet).

Obscured Figures Test (OFT). The OFT was designed to assess speed of closure, defined as ". . . the degree to which different pieces of information can be combined and organized into one meaningful pattern quickly. It is not known beforehand what the pattern will be." The OFT was originally based on the "Gestalt Completion Test" (GCT) in the ETS kit, but it was altered substantially to conform to a multiple-choice response format. The GCT was included in the experimental battery.

The OFT is divided into two, individually timed sections containing 28 questions each (see Appendix I). On each question, a partially obscured target figure is presented and followed by four alternative figures from the same class (e.g., flowers, trees, fruit). The subject must determine which of the alternative figures is identical to the obscured target figure. In the "White" section, the obscured targets are black figures presented on a white background. In the "Black" section, the obscured targets are white figures presented on a black background. The target figures are otherwise identical in the two sections, although they are presented in a different order. The order of presentation of the sections was counterbalanced across subjects($n_1$ = 135; $n_2$ = 138).

The OFT and GCT use figures as stimulus materials in an attempt to measure the same ability construct. The tests are otherwise very different. The two sections of the GCT are equivalent in format (always an obscured, dark figure on a white background) but different figures are presented in each section. In contrast, both sections of the OFT use the same figures but present them in a different format. The GCT is an open-ended response format as opposed to a multiple-choice format for the OFT. The tests also differ in the number of items (56 versus 40) and the answer format (marking a separate answer sheet versus marking on the test sheet). The OFT is, in many ways, more similar to the RMT than to the GCT.

Preliminary Test Development Procedures

The nine new ability tests that compose the experimental NFAST battery were generally developed in the order in which the tests were described above. The rationale for the test development order was to begin with tests that were expected to require the most extensive pretesting and revision. As tests were developed, they were administered to small samples of subjects and then revised while other tests were being developed. This procedure optimized the availability of subject personnel during the preliminary test development process. An overview of the preliminary testing process is described below.

Since the complex, cognitive-perceptual test was not modeled after another test or developed from other specifications, detailed test specifications were developed first (see Appendix J) and an initial version of the test was then produced. The first version of the Flight Planning Test (FPT) was administered to two civilians and five experienced helicopter pilots. The results of the preliminary test administration were positive: the subjects required less time to complete the items than had been expected, the test scores were indicative of substantial individual differences, the subjects reported that the FPT had a high degree of relevance to helicopter performance (i.e., high "face validity"), and no serious difficulties were encountered with the instructions or test procedures.

The FPT was subsequently administered to a panel of seven behavioral scientists and one aviation officer as part of a peer review of the project. The panel members were asked to critique the test instructions, items, time limits, and concept, and to make suggestions for improving the test. Again, the reactions to the FPT were generally positive, although several constructive modifications were recommended. The recommendations were primarily directed toward reducing the level of difficulty of the test. For example, it was recommended that additional instructions and practice items should be provided, especially on items that required mathematical computations.

The peer review recommendations were evaluated and a plan for revising the test was developed. As the FPT revisions were being implemented, the Chart Use Test (CUT), Map Planning Test (MPT), Finding Rules Test (FRT), and Sound Reasoning Test (SRT) were being developed. A large number of items were written for each test and then reviewed to select the items that were expected to assess the desired range of ability levels (i.e., items that appeared to range from very easy to very difficult). The battery of five tests was administered to 14 Warrant Officer Candidates (WOCs) and 13 second lieutenants (2LTs) who were awaiting the convening of their flight training classes. Test administration time for the five tests was approximately five and one-half hours, including breaks and time to complete a standardized critique.

The reactions of the subjects to the tests were, again, generally positive. The FRT received the most unfavorable reactions. Specifically, it was perceived to be excessively difficult and not relevant to aviation training. Although the SRT is also relatively abstract, its difficulty level, time limits, and perceived relevance to aviation training were acceptable to the subjects. The CUT was rated as being highly relevant and having an acceptable difficulty level, but the time limits imposed were perceived to be much too short. Reactions to the FPT and MPT were positive.

The information obtained from the last administration was used to make further revisions to the initial five tests. In addition, changes were made to standardize the tests into a common format. The last four tests, modeled after the ETS kit tests, were developed and produced for preliminary administration. The entire battery of nine tests was administered to three separate groups ($n$ = 10, 11, and 11, respectively) of flight students awaiting the convening of their classes. The total test time was approximately 7 hours. Different time limits and orders-of-presentation were used for each administration. Between each administration, minor revisions were made to the tests, as needed. The nine preliminary tests, their sections, and their items were analyzed and then revised for administration as an experimental battery. Final revisions were also made to all ancillary materials and to the standardized test protocol.

Experimental Battery Administration Procedures

The experimental battery was administered to six groups of subjects located at three military bases in the southeastern United States. The group sizes, in order of administration, were 33, 37, 46, 155, 10, and 9. Each administration followed the same protocol except for the order of presentation of the tests. The experimental battery was administered using the three orders-of-presentation shown in Table 2. Eleven of the tests were clustered into three groups: the ETS group (CRT,HPT, IPT, and GCT); the ETS-modeled group (FOT, FFT, RMT, and OFT); and the Hills, et al. (1983) group (SRT, MPT, and FRT). The tests in each group were always presented in the order shown above or in the reverse order. Furthermore, the first and second groups (the latter modeled after the former) were always presented in separate halves of the test day (morning or afternoon). Most importantly, the FPT was presented at various times during the middle of the test day (mid morning, late morning, or early afternoon). The placement of the FPT in the battery was intended to minimize warm-up or fatigue effects due to time-of-day and learning effects from the other tests in the battery.

As indicated in the test descriptions, the order of presentation of the developmental test sections was completely counterbalanced within the constraints of the respective test designs (i.e., increasing the level of difficulty on the FPT and CUT). In addition, the combination of presentation series across

Table 2

Orders of Presentation of the Experimental Battery

| Order 1 | Order 2 | Order 3 |
|---|---|---|
| CRT | FOT | OFT |
| HPT | FFT | RMT |
| IPT | RMT | FFT |
| GCT | OFT | FOT |
| CUT | SRT | FPT |
| FPT | MPT | (LUNCH) |
| (LUNCH) | FRT | FRT |
| FOT | CUT | MPT |
| FFT | (LUNCH) | SRT |
| RMT | FPT | CUT |
| OFT | CRT | GCT |
| SRT | HPT | IPT |
| MPT | IPT | HPT |
| FRT | GCT | CRT |

Note. Order 1 (n = 79) was used for administrations 1 and 3 ; order 2 (n = 56) was used for administrations 2, 5, and 6; and order 3 (n = 155) was used for administration 4. FPT = Flight Planning Test; CUT = Chart Use Test; MPT = Map Planning Test; SRT = Sound Reasoning Test; FRT = Finding Rules Test; RMT = Rapid Match Test; OFT = Obscured Figures Test; FFT = Finding Figures Test; FOT = Figure Orientation Test; IPT = Identical Pictures Test; GCT = Gestalt Completion Test; HPT = Hidden Patterns Test; CRT = Card Rotations Test.

the tests was partially counterbalanced. For example, the test packets were prepared so that the subjects who took the FPT in the first order-of-presentation series would take each of the MPT series in approximately equal numbers. This procedure was extended to the other FPT series and to the other tests in the battery. That is, approximately half of the subjects who took the first FPT series took the FOT in the rotated-inverted order and half took the FOT in the opposite order. Sixty series combinations were selected to control for order effects across the tests. Five packets were prepared for each of the combinations. The packets were randomly distributed during each administration.

Test administrations began at approximately 8:00 a.m. and concluded at approximately 4:00 p.m. The subjects were given an hour lunch break and liberal rest breaks between tests to minimize the effects of fatigue. Each administration was monitored by one to four proctors who were permitted to answer only procedural questions during the test periods. Test instructions were read to the subjects and questions were entertained before beginning each test. Each test section was separately timed. At the end of the test day, the subjects were asked to rate the level of effort they had expended and the percentage of item responses that had been guesses. Prior to dismissal, the subjects were thanked for their participation and provided an opportunity to ask questions about the tests and their interpretation, as well as questions about aviation selection and training. Subjects who requested feedback on their performance were subsequently sent a summary of their test scores, normative data on the entire sample, and instructions for interpreting their performance.

Following each administration, the base Adjutant General's office was asked to provide the ASVAB General Technical (GT) composite score from each subject's personnel record.

Subjects

Data were collected from 290 subjects during the six administrations of the experimental battery. The subjects were recruited at each base by a local point-of-contact. Since the experimental battery was always administered during the duty day, the subjects were, in effect, paid for their participation. However, no remuneration other than regular salary was paid. The incentives that were offered to recruit the subjects included the opportunity to assist in developing the battery, a change from their regular duties, the opportunity to develop their test-taking skills, and the opportunity to compete with their peers in a nonthreatening situation (subjects who requested feedback were furnished with their test scores, normative data, and instructions for interpreting their scores). The subjects were treated in accordance with the "Ethical Principles of Psychologists" (American Psychological Association, 1981).

Data from five subjects were deleted for failing to complete substantial portions of the experimental battery. The subjects either did not return from the lunch break or were excused for medical or personal reasons. Data from one subject was deleted due to prior exposure to the preliminary test materials (i.e., he had served as a subject during preliminary testing).

Complete data were collected from the remaining 284 subjects. However, 36 subjects were identified as possibly violating test procedures or not giving a reasonable effort. The subjects were identified either during the administration of the battery (e.g., not attempting sections of a test, becoming drowsy, or turning back to test instructions in violation of standard procedures) or from self reports (subjects rated their level of effort and percentage of items guessed after completing the battery). The data files of the 36 subjects were evaluated for possible elimination from further analyses. The primary decision rule was that the preponderance of evidence must be against the retention of the subject's file before the data were eliminated. The following evidence was considered in determining whether to retain or eliminate a subject from the sample:

- self reports of effort level, percentage of items guessed, and attitude ratings on the biographic information sheet,
- notations of suspicious behavior during the administration of the battery,
- not answering any items on a test section without explanation,
- answering all the items on highly speeded tests,
- having test scores that were at or below chance levels,
- exhibiting a suspicious or repetitive answer pattern,
- use of scratch paper and notes on test materials, and
- consistency within and across tests.

Each of the suspect subject files was independently evaluated by two research psychologists. The data from 11 (3.8 percent of the original sample) subjects were eliminated from further analyses. The data from several other subjects were rated as very marginal (i.e., one of the two evaluators recommended elimination), but the data were retained in accordance with the conservative decision rule. All of the subsequent analyses are based on the remaining 273 subjects, unless otherwise indicated. Further biographical and attitudinal information about the subjects is reported in the Results section.

# Results and Discussion

The results of this research are presented in three parts. Biographic and attitudinal information about the subjects is presented in the first part. The subject information was obtained from the Biographic Information Sheet completed by the 273 subjects that are included in the analyses. Descriptive statistics on each of the tests in the experimental battery are presented in the second part. More detailed statistics on the sections of the Flight Planning Test (FPT) are also presented in this part. Finally, the factor analytic results are presented in the third part.

## Subject Information

Virtually all the subjects were active duty military personnel (see Table 3). Of the two nonmilitary subjects, one was a civilian government employee and one was an ROTC cadet in summer training. The military subjects represented 33 U.S. Army occupational specialties, but only 19 subjects reported having any aviation maintenance or flight training experience. Only two subjects had more than 50 flight hours. The majority (63 percent) of the military subjects had less than one year of service and 90 percent of the subjects had 8 years or less of service.

The ages of the subjects ranged from 17 to 44 years old. However, the median age was 21 years and 85 percent of the subjects were between 17 and 28 years. The age distribution of the sample closely approximates the age distribution of the population of Army enlisted personnel (Department of Defense, 1985). The majority (65 percent) of the subjects had never been married (see Table 3). In contrast, approximately 50 percent of all Army enlisted personnel are currently married (Department of Defense, 1985). None of the subjects reported having any serious visual problems.

Table 3

Subject Demographic Information

| Position | Number |
|---|---|
| Civilian | 2 |
| Enlisted Service Member | 171 |
| Noncommissioned Officer | 78 |
| Warrant Officer Candidate | 10 |
| Commisioned Officer | 12 |
| **Marital Status** | |
| Never Married | 178 |
| Married | 68 |
| Separated or Divorced | 18 |
| No response | 9 |
| **Sex** | |
| Male | 215 |
| Female | 58 |
| **Ethnic Background** | |
| White | 203 |
| Black | 55 |
| Hispanic | 8 |
| Asian | 3 |
| Other or no response | 4 |
| **Educational Degrees** | |
| Less than High School | 31 |
| High School | 183 |
| Two-year or Vocational College | 38 |
| Four-year College | 21 |

Table 3 also presents the sex and ethnic background distribution of the sample. The percentage of female (21 percent) subjects is slightly higher than in the Army population (10 percent) and the percentage of minority (26 percent) subjects is slightly lower than the percentage in the Army population (35 percent; Department of Defense, 1985). In addition, the two categories overlap considerably: twenty-four of the subjects were both female and from an ethnic minority.

The number of years of formal education ranged from 9 to 17 years, but the majority (54 percent) reported 12 years of education. The highest educational degree awarded closely parallels the number of years of education (see Table 3). There are several significant relationships ($\chi^2$, $p<.05$) between the reported educational levels and degrees attained, and the sex and ethnic backgrounds of the subjects. Women and members of an ethnic minority were more likely than white men to report having between 13 to 15 years of formal education and to hold a two-year or vocational/technical degree. White men were more likely to report having less than 12 years of formal education and not holding at least a high school degree.

However, there was contradictory information contained in the Armed Services Vocational Aptitude Battery (ASVAB) General Technical (GT) composite scores. The ASVAB GT, composed of word knowledge, paragraph reasoning, and arithmetic reasoning aptitude scores, is commonly considered to be a general mental ability or academic aptitude test (e.g., Maier & Grafton, 1981; Rossmeissl, Martin, & Wing, 1983). Despite their greater educational achievements, on the average, women and ethnic minority members scored substantially lower on the GT composite than white men.

The original research plan provided for analyses of performance on each test as a function of subgroup membership and educational level. However, the substantial overlap between the female and minority categories in the sample confounds attributions to either demographic category. In addition, the contradictions between the self-reported educational achievements and the confirmed scores on a nationally normed, academic aptitude test indicate that the self reports may be unreliable. Although the GT scores are not completely reliable (e.g., unreliability due to the length of time since the ASVAB was administered), internal inconsistencies indicate the self-reported information is even less reliable. For example, some subjects reported less years of education than would be required for the claimed degree. In addition, approximately 19 percent of the subjects who claimed only a high school degree indicated they had more than 12 years of education. Whether those additional years were based on calendar years of study or academic credits is unknown. As a consequence of these confounds and contradictions, further subgroup analyses were not attempted. The GT scores were included in the analyses as a measure of general mental ability or academic aptitude.

The subjects also responded to three attitudinal questions before taking the experimental battery. First, 32 percent of the subjects indicated they had no interest in aviation while 40 percent indicated they had a strong interest in aviation of any type. The remaining subjects were interested in rotary-wing aviation only (25 percent) or fixed-wing aviation only (3 percent). Second, the subjects indicated how well they expected to perform on the battery of tests: well below average (2 percent), below average (9 percent), average (46 percent), above average (36 percent), and well above average (7 percent). Finally, 1 percent of the subjects indicated they disliked taking the test battery, 9 percent preferred to work at some other task, 10 percent preferred taking the battery to working at some other task, and 80 percent indicated they were glad to help in developing the tests.

Descriptive Statistics

Several dependent measures were calculated for the tests in the experimental battery. A dichotomous score of correct-incorrect was, of course, used for the analysis of test items. The primary measure of static ability was the number of questions answered correctly (NCORR). NCORR scores were computed for each test and for each test section. Scores were also computed for each test to evaluate the effects of guessing. Each corrected-for-guessing (CFG) score was calculated by subtracting one-third of the number of incorrectly answered items from the NCORR (e.g., Cronbach, 1970, pp. 56-58). Finally, the percentage of attempted items that were answered correctly (%ACC) was computed for each test to assess the relative importance of speed versus accuracy.

Six additional measures were derived for each subject from their FPT section scores. First, the NCORR for each section of the FPT was regressed on the level of difficulty (i.e., levels 1, 2, and 3). The intercept (INT) of the regression equations was analyzed as an index of performance under minimum information demands. The slope (SLP) of the regression equations was analyzed to assess the effects of increasing the amount of information to be processed.

Second, two gain scores were computed for each of the second (GS1 and GS2) and third (GS3 and GS4)levels of difficulty. The first section of each level of difficulty was considered a baseline measure that was subtracted from the NCORR for each of the next two sections. Twelve points were added so that all gain scores were positive values. The gain scores were analyzed as indices of the ability to learn as a function of review and practice at different levels of difficulty. Gain scores 1 and 3 represent one practice and review iteration and gain scores 2 and 4 represent two practice and review iterations.

Number of correct responses. As shown in Table 4, the average NCORR for the nine developmental tests ranges from 39.2 to 62.0 percent of the maximum possible score. All of the tests were designed to have an average item difficulty of approximately .50. However, the FPT and the CUT were also designed to impose an increasing level of difficulty as the test progressed, resulting in the lower NCORR means. The higher RMT mean is primarily a function of the time limits imposed. The distribution of scores for each test was approximately symmetrical; none of the skewness indices were significant. More importantly, each of the developmental tests exhibits a relatively high degree of variability, indicating the measurement of substantial individual differences. The coefficients of variability (the standard deviation divided by the mean) range from 20.4 to 43.8 percent.

The mean and standard deviation of the sections of each developmental test were generally very similar, although there were significant mean differences among the sections on every test ($p < .01$). As noted above, the sections of the FPT and CUT were designed to present increasing levels of difficulty; these differences are reflected in the descriptive statistics on their sections (see the FPT section analysis). The two FFT sections had substantially different means (24.5 versus 35.5; $t(272) = 17.35$, $p < .001$). As expected, the geometric figure that could be embedded in one of three positions within the pattern produced much lower average scores than the figure that could only be embedded in the center of the pattern. The remaining differences reflect the minor variations that were designed into the test sections (e.g., finding the identical figure versus finding the different figure on the RMT; cf. Farell, 1985).

In addition to the statistics based on the identity of the sections (e.g., the inverted and rotated sections of the FOT, the black and white background sections of the OFT), descriptive statistics based on the order of presentation of

Table 4

Descriptive Statistics on the Experimental Battery Number Correct (NCORR) Scores (N = 273)

| Test[a] | Sections | Items | Mean | SD | Alpha[b] | Section r[c] |
|---|---|---|---|---|---|---|
| FPT | 7 | 84 | 34.6 | 11.6 | .88 | .43 |
| CUT | 5 | 30 | 11.8 | 5.0 | .80 | .40 |
| MPT | 3 | 60 | 31.4 | 10.3 | .93 | .60 |
| SRT | 2 | 32 | 16.2 | 5.1 | .82 | .69 |
| FRT | 2 | 32 | 14.0 | 5.5 | .79 | .62 |
| RMT | 2 | 96 | 59.5 | 12.1 | .94 | .62 |
| FOT | 2 | 84 | 39.0 | 17.1 | .96 | .71 |
| FFT | 2 | 120 | 60.0 | 24.8 | .97 | .71 |
| OFT | 2 | 56 | 31.7 | 8.7 | .87 | .38 |
| IPT | 2 | 96 | 74.7 | 13.7 | - | .73 |
| CRT | 2 | 160 | 118.8 | 25.2 | - | .75 |
| HPT | 2 | 400 | 198.1 | 45.7 | - | .80 |
| GCT | 2 | 20 | 14.2 | 3.1 | - | .46 |
| GT[d] | - | - | 111.1 | 11.4 | - | - |

[a] FPT = Flight Planning Test; CUT = Chart Use Test; MPT = Map Planning Test; SRT = Sound Reasoning Test; FRT = Finding Rules Test; RMT = Rapid Match Test; FOT = Figure Orientation Test; FFT = Finding Figures Test; OFT = Obscured Figures Test; IPT = Identical Pictures Test; CRT = Card Rotations Test; HPT = Hidden Patterns Test; GCT = Gestalt Completion Test; GT = General Technical score from the Armed Services Vocational Aptitude Battery.

[b]Coefficient alpha for the total test; not computed for the ETS tests or the GT.

[c]Correlation between the sections of each test; if the test had more than two sections, the mean correlation is reported.

[d]Only the total GT score from the ASVAB was available ($n$ = 245).

the sections were also computed . With the exception of the CUT (which was presented in a fixed order) and the FPT (which will be discussed later), practice on the first section of each developmental test resulted in a consistent improvement, on the average, in performance on the subsequent section(s). The magnitude of the order differences was approximately the same as the differences between sections based on identity. All of the differences were statistically significant ($p < .01$ except the FRT, $p < .05$).

Table 4 also presents two estimates of the reliability of each test. The coefficient alphas are acceptably high (see Nunnally, 1978, p. 278), indicating reasonable levels of internal consistency. The coefficient alphas on the highly speeded tests are probably inflated. The coefficients for the test sections are lower than the total test, but they are in proportion to the number of items in each section (see Cronbach, 1970, pp. 165-171). The last column presents the correlation (or mean correlation, as appropriate) between sections of each test as an estimate of their equivalence. The estimates of reliability are modest but acceptable when corrected for test length; the Spearman-Brown corrected coefficients range from .55 to .83. The OFT is the only test with an unacceptable correlation between the sections. The FPT and CUT also have relatively low reliability estimates, but they were not designed to have equivalent sections.

Descriptive statistics on the ETS and ASVAB GT scores are also presented in Table 4. The ETS tests have relatively higher mean scores (49.5 to 77.9 percent of the maximum score possible) and exhibit relatively lower variability (the coefficients of variation range from 18.3 to 23.1) than the developmental tests. However, the means and standard deviations are very similar to the results of a previous administration to 275 WOCs and 2LTs awaiting flight training (Myers, Schemmer, & Fleishman, 1983). On the four ETS tests, the current sample of subjects is approximately equivalent to a sample of entering flight students.

Since the ETS tests were hand scored and only the NCORR and number attempted for each part were entered in the data base, coefficient alphas were not computed. The correlations between the two parts of the ETS tests are slightly higher, except for the GCT, than the developmental tests. However, the ETS tests were designed to be equivalent forms (see the test descriptions) while the ETS-modeled tests were designed to have some differences between the sections.

The ASVAB GT composite score was obtained from official records for 245 of the 278 subjects. Since the GT is a standardized score derived from three independent ability test scores (not available), neither estimate of reliability could be calculated. There are two noteworthy points concerning the mean and standard deviation, however. First, the average GT score is only slightly higher than the minimum (110) required for acceptance as an IERW student. Despite similarities to recent flight students on the ETS tests, nearly half of the subjects would not qualify for flight training on the basis of their GT scores. Second, the variability of the GT score is relatively low (10.3 percent of the mean) despite the wide range of subject backgrounds. The lack of variability limits the maximum correlation between the GT and other scores in the battery.

NCORR intercorrelation matrix. The NCORR intercorrelation matrix of the nine developmental tests, the four ETS tests, and the ASVAB GT is presented in Table 5. The FPT is most highly correlated with the CUT and the MPT. The CUT is also a relatively complex cognitive-perceptual test that involves the opportunity to improve one's performance with experience, although the sections are not cumulative as they are in the FPT. As was mentioned in the test descriptions, the MPT is at least superficially a less difficult version of the

Table 5

Intercorrelation Matrix of NCORR Scores

| | FPT | CUT | MPT | SRT | FRT | RMT | FOT | FFT | OFT | IPT | CRT | HPT | GCT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CUT | 60 | | | | | | | | | | | | |
| MPT | 56 | 58 | | | | | | | | | | | |
| SRT | 30 | 42 | 43 | | | | | | | | | | |
| FRT | 40 | 41 | 44 | 36 | | | | | | | | | |
| RMT | 46 | 45 | 55 | 35 | 43 | | | | | | | | |
| FOT | 44 | 51 | 49 | 42 | 40 | 47 | | | | | | | |
| FFT | 48 | 45 | 53 | 34 | 44 | 52 | 54 | | | | | | |
| OFT | 45 | 38 | 48 | 28 | 37 | 59 | 39 | 38 | | | | | |
| IPT | 36 | 36 | 47 | 31 | 35 | 57 | 38 | 45 | 56 | | | | |
| CRT | 37 | 41 | 51 | 32 | 32 | 39 | 65 | 44 | 39 | 48 | | | |
| HPT | 48 | 46 | 54 | 32 | 33 | 52 | 46 | 62 | 43 | 58 | 51 | | |
| GCT | 07 | 22 | 17 | 15 | 14 | 12 | 16 | 28 | 21 | 22 | 25 | 16 | |
| GT | 41 | 56 | 41 | 40 | 33 | 29 | 34 | 41 | 22 | 22 | 30 | 42 | 26 |

Note. Decimals have been omitted.

FPT. The CUT and MPT are also highly correlated with each other. The FPT is least correlated with the GCT. However, the GCT has only low correlations with all the tests in the battery and a relatively low correlation (.46) between the two sections of the test. The low correlations can be partially attributed to the low degree of variability of the GCT scores. It is also the only test in the battery that has an open-ended response format and that is subjectively scored.

The ETS tests, with the exception of the GCT, are more highly correlated with their corresponding developmental tests than with any other test in the battery. The OFT is most highly correlated with the RMT and the IPT. The two measures of reasoning (SRT and FRT) and of closure (FFT and OFT), while conceptually similar, do not share much common variance.

FPT section analyses. The mean NCORR scores for the sections of the FPT (see Table 6) indicate that the design specifications for the levels of difficulty were generally met, especially for the identity sections. There are significant differences between the sections of the FPT ($F(6, 1632) = 63.17$, $p < .001$), primarily between sections at different levels of difficulty (Tukey (a), $p < .01$). Performance was consistently lower as the level of difficulty increased. There were no significant differences between the sections at the second level of difficulty or between two of the sections (3E and 3W) at the third level of difficulty. Performance on section 3N was more similar to the second level of difficulty.

The average number of items attempted for section 1 is 11.2 out of a maximum of 12. Section 1 appears to be relatively unaffected by the current time limits. The time constraints were a larger factor at the second and third levels of difficulty. The average number of items attempted on the other sections ranges from 8.5 to 8.9. The variability of test scores is both homogeneous between sections and indicative of substantial individual

Table 6

Descriptive Statistics for the FPT Sections Based on Item Identity and Order of Presentation

| | Section Identity | | | Section Order | | |
|---|---|---|---|---|---|---|
| Difficulty | Section[a] | Mean | SD | Section[b] | Mean | SD |
| Level 1 | FPT1S | 6.28 | 2.26 | FPT1 | 6.28 | 2.26 |
| Level 2 | FPT2E | 5.49 | 2.47 | FPT2 | 4.70 | 2.25 |
| Level 2 | FPT2N | 5.10 | 2.71 | FPT3 | 5.18 | 2.38 |
| Level 2 | FPT2W | 5.13 | 2.18 | FPT4 | 5.84 | 2.64 |
| Level 3 | FPT3E | 4.02 | 2.08 | FPT5 | 4.41 | 2.26 |
| Level 3 | FPT3N | 4.82 | 2.33 | FPT6 | 4.08 | 2.19 |
| Level 3 | FPT3W | 3.80 | 2.15 | FPT7 | 4.15 | 2.23 |

Note. The levels of difficulty were always presented in the same order (1, 2, and 3, respectively). The order of presentation of the identity sections was completely counterbalanced within each level of difficulty. In addition, the three presentation series in difficulty levels 2 and 3 were fully crossed, resulting in nine orders of presentation.

[a] The identity sections are designated by the test acronym, the difficulty level number, and an arbitrary letter disignation that implies a map location (South, East, North, or West).

[b] The order sections are designated by the test acronym and a number representing the order of presentation.

differences within sections.

The order of presentation of the sections also had a significant effect on the NCORR means ($F$ (6, 1632) = 63.83, $p < .001$). After a relatively large decrement (25 percent) in performance from level 1 to the first section of level 2, the mean score for the next two sections steadily increases to 93 percent of the level 1 value. The NCORR means for the level 1 and level 2 sections are significantly different from each other (Tukey (a), $p < .01$). Performance at difficulty level 3 does not exhibit the same pattern. The mean score for section 5 is 24 percent lower than the mean score for section 4. However, performance on sections 6 and 7 are stable and slightly (6 percent) lower than performance on the initial section of level 3. The level 3 sections are significantly different (Tukey (a), $p < .01$) from the other levels of difficulty (except sections 2 and 5), but are not different from each other. Performance at the third level is difficult to interpret because of the lack of any indication of learning. Performance based on the order of presentation may also be confounded by unintended differences in the difficulty of the three level 3 sections (see the identity statistics in Table 6).

Item analysis of the developmental tests. Each of the developmental tests was evaluated in terms of four, standard psychometric criteria (cf. Nunnally, 1978, pp. 261-285). The average item difficulty (the percentage of correctly answered items) for each test should be approximately .50 to maximize the potential variability of scores. The difficulty level of the individual items should vary about the average value, but no item should be answered correctly or incorrectly by all the subjects. Finally, there are two important aspects to the item-total correlations. First, the correlations should be positive in sign; subjects who obtain higher scores on the test should have a higher probability of getting each item correct. Second, a substantial percentage of the correlations should be .30 or higher if the test is to be considered homogenous.

Table 7 presents the descriptive statistics on item difficulty and the corrected item-total correlations for the nine developmental tests in the battery. The average item difficulty for each test is near the optimum level of .50. The RMT and the CUT have the most extreme item difficulties. The average item difficulties are, of course, reflected in the mean NCORR as well. More importantly, the items on each test exhibit a wide range of difficulty levels. The RMT has the largest range (.98) while the FRT has the lowest, but still acceptable, range (.60). None of the items are always answered correctly or incorrectly, although the MPT, RMT, FOT, and FFT have items that approach these extremes.

The corrected item-total correlations also exhibit desirable psychometric characteristics. First, there are relatively few negative correlations and only one (-.25) is of a substantial magnitude (the item is not miskeyed; it is very difficult ($p$ = .08) and the correlation may be substantially affected by guessing). Second, a large proportion of the items on each test have an acceptably high (i.e., $r \geq$ .30) correlation with the total score. The FPT has the lowest proportion of highly correlated items. This is not surprising since the total FPT is a measure of a complex of abilities; each item assesses only a portion of the abilities. Each of the six types of questions are represented at least three times (landmark and airspeed questions) but not more than eight times (flight time) in the 29 items that are most highly correlated with the FPT NCORR. Eleven of the 29 questions contain a selective attention distractor. That is, 45.8 percent of the questions that contained a selective attention distractor correlated at least .30 with the total FPT NCORR.

%ACC and CFG descriptive statistics. The descriptive statistics for the %ACC and CFG measures (see Table 8) are predictable from the NCORR results. The %ACC means readily divide into two categories: power tests that

Table 7

Item Analysis Statistics for the Nine Developmental Tests

| | Item difficulty | | | | Item-total correlations | | | |
|---|---|---|---|---|---|---|---|---|
| Test | Mean | SD | Min | Max | Min | Max | %neg[a] | %>.3[b] |
| FPT | .41 | .18 | .08 | .81 | -.03 | .53 | .01 | .35 |
| CUT | .39 | .23 | .03 | .75 | -.15 | .51 | .07 | .60 |
| MPT | .53 | .34 | .02 | .97 | .02 | .66 | .00 | .75 |
| SRT | .51 | .29 | .04 | .86 | -.25 | .60 | .16 | .66 |
| FRT | .44 | .16 | .15 | .75 | .05 | .52 | .00 | .47 |
| RMT | .62 | .36 | .01 | .99 | -.05 | .69 | .02 | .51 |
| FOT | .47 | .29 | .02 | .85 | .15 | .69 | .00 | .90 |
| FFT | .50 | .32 | .01 | .96 | .31 | .78 | .00 | 1.00 |
| OFT | .57 | .28 | .08 | .98 | -.10 | .60 | .05 | .61 |

Note. Min = minimum; max = maximum; items that were always answered incorrectly would have a difficulty of .00 and items that were always answered correctly would have a difficulty of 1.00; all item-total correlations are corrected for autocorrelation.

[a] %neg = proportion of negative item-total correlation coefficients.

[b] % >.3 = proportion of item-total correlation coefficients of .30 or greater.

have an accuracy level of approximately 55 percent and speed tests that have an accuracy level greater than 80 percent. The average difficulty level of power tests is manipulated by the difficulty of the items since the time limits are sufficient to attempt most of the items. The average difficulty level of speed tests is manipulated by the use of restrictive time limits since the items are of trivial difficulty. As a result, tests with a high accuracy rate are considered speed tests while tests with a moderate accuracy rate are considered power tests. The two types of tests tend to measure different factors despite similarities in their construction (cf. Nunnally, 1978, pp. 629-639).

The ETS tests, with the exception of the GCT, are the most highly speeded tests in the battery. The RMT is the most highly speeded developmental test. The FPT, CUT, SRT, and FRT are all power tests with an average accuracy level near .55. The variability of the %ACC scores is primarily a function of the speed-accuracy dichotomy. Compared to the NCORR measures (see Table 4), the coefficients of variability for the power tests are approximately the same while the coefficients for the speed tests are substantially lower.

As expected, changes in the mean CFG scores (see Table 8) are also a function of the level of accuracy: the means of the power tests are substantially lower than the NCORR means while the means of the speeded tests are essentially unchanged. Only the least capable subjects need to guess on the speed tests since all the items are of minimal difficulty. On the power tests, subjects at all levels of ability may need to guess at some of the more difficult items. The CFG score is intended to assess individual differences in the tendency to guess (Nunnally, 1978, pp. 647-650). For all the tests, the coefficient of variability is higher than the corresponding NCORR coefficients. The increase in the relative variability of the CFG scores is most dramatic for

Table 8

Descriptive Statistics for the Percent Accuracy (%ACC) and Corrected-for-Guessing (CFG) Measures.

| Test | % ACC mean | %ACC SD | CFG mean | CFG SD |
|---|---|---|---|---|
| FPT | 54.63 | 13.97 | 25.13 | 13.11 |
| CUT | 55.86 | 20.51 | 8.79 | 5.98 |
| MPT | 88.23 | 15.80 | 30.23 | 10.88 |
| SRT | 57.03 | 14.97 | 12.21 | 5.73 |
| FRT | 53.85 | 19.48 | 10.02 | 6.37 |
| RMT | 91.87 | 8.08 | 57.75 | 12.65 |
| FOT | 78.29 | 21.95 | 35.84 | 18.80 |
| FFT | 87.30 | 15.49 | 57.74 | 25.97 |
| OFT | 83.54 | 9.98 | 29.65 | 8.89 |
| IPT | 96.54 | 2.58 | 74.08 | 13.67 |
| CRT | 91.88 | 8.32 | 108.56 | 30.27 |
| HPT | 96.04 | 4.96 | 190.64 | 48.71 |
| GCT | 83.98 | 13.74 | 13.31 | 3.42 |

the power tests because the means are lower and the standard deviations are higher than the corresponding NCORR statistics.

Although the group averages for the %ACC and the CFG can be computed directly from the NCORR scores if the number of items attempted is known, the assessment of individual differences in tendencies toward speededness or accuracy in responding and in predisposition toward guessing may alter the NCORR factor structure of the experimental battery.

FPT regression and gain score statistics. The final two sets of dependent measures were the intercept (INT) and slope (SLP) from the regression of FPT section NCORR scores on the level of difficulty and the gain scores (GS1-GS4) for difficulty levels 2 and 3. The mean (and standard deviation) of the INT and SLP is 7.30 (2.85) and -1.03 (1.02), respectively. As would be expected from the analysis of the FPT sections, increasing the level of difficulty results in an average NCORR decrease of approximately one point. There are substantial individual differences apparent in both measures.

The mean (and standard deviation) of the gain scores is: GS1 = 12.48 (2.42) and GS2 = 13.14 (2.37) for difficulty level 2; GS3 = 11.67 (2.52) and GS4 = 11.74 (2.36) for difficulty level 3. The gain scores for difficulty level 2 reflect the slight learning effect observed in the FPT section analysis while the difficulty level 3 gain scores reflect a slight decrement in performance A gain score of 12 would indicate no change at all. The standard deviations indicate that the derivation of the gain scores reduces the relative degree of variability. That is, a large part of the individual differences in within-difficulty-level performance is common to each of the three sections. However, additional variance is accounted for by the second and third sections of each difficulty level.

The means for each of the regression and gain score measures could have been computed directly from the FPT section analyses. They are presented as a reference to evaluate the standard deviations. As was noted with regard to the %ACC and CFG data, it is the individual differences in these measures rather than the group averages that are most important in the evaluation of the FPT. The different measures that have been derived from the FPT NCORR scores may contribute independent sources of variance to the underlying factor structure of the experimental battery.

Factor Analyses

A series of maximum-likelihood factor analyses (BMDP4M; Dixon, et al., 1983) with varimax rotation were used to analyze the factor structure of the experimental battery. The data were input as a correlation matrix (BMDPAM) to enable a pairwise deletion of cases that were missing GT scores. The purpose of these analyses was to condense each data matrix to a smaller number of underlying common factors. The analyses were used to interpret the ability constructs or domains that are being measured by the battery, to determine if the factor structure is a function of different characteristics of the tests, and subsequently to make decisions concerning further analyses or research.

The analyses were conducted in three phases. In the first phase, two analyses were conducted to determine the factor structure of the battery based on the identity and the order of presentation of the test sections. In the second phase, three analyses were conducted to determine if the NCORR, CFG, and %ACC measures resulted in different factor structures. In the final phase, the FPT regression and gain score measures were included in the NCORR factor analysis to determine if they assess independent dimensions of performance.

Only loadings of .40 or greater are shown on each table of factor analytic results. Factor loadings of less than .40 are omitted both in the interest of clarity and to avoid the overinterpretation of small loadings (e.g., Comrey, 1973; Nunnally, 1978). For example, a small loading may be given undue weight if it is the highest loading that the variable has on any factor. In lieu of the complete table of factor loadings, the communality ($h^2$) of each test is presented to show the variance that each test has in common with the factors.

Section factor analyses. The initial factor analysis was conducted to determine the factor structure of the experimental battery scores based on the identity of the test sections. A seven-factor solution was obtained that accounted for 52.6 percent of the variance (see Table 9). The first three factors account for 32.1 percent of the variance in approximately equal proportions. The first factor is defined by the IPT and RMT sections, both measures of perceptual speed. Both sections of the OFT also load on this factor. As has been noted previously, the OFT is similar to the RMT except that the target figures are obscured. The factor clearly appears to represent a perceptual speed (PSPD) ability.

The second factor is defined by the five CUT sections and the ASVAB GT. Both tests require the use of verbal and quantitative abilities. An individual's performance may also be influenced by prior knowledge on either test, although the CUT sections were designed to minimize the effects of technical expertise. Most importantly, both tests require the subject to comprehend and apply explicit rules to produce a correct response, although there are drastic differences in test content. The similarities are greatest between the CUT and the paragraph comprehension portions of the GT composite. With some trepidation, the second factor is labeled as an information ordering (INOR) ability.

The third factor is defined by six of the seven sections of the FPT and is accordingly interpreted as a complex, cognitive-perceptual (CGPR) ability. The emphasis is placed on the complexity of performance since the first level of difficulty (defined in terms of complexity of information to be processed) does

Table 9

Factor Analysis of the NCORR Scores Based on Section Identity

| Section | PSPD | INOR | CGPR | FCLO | SPOR | DEDR | INDR | $h^2$ |
|---|---|---|---|---|---|---|---|---|
| FPT1S | | | | | | | | 37 |
| FPT2E | | | 58 | | | | | 50 |
| FPT2N | | | 64 | | | | | 51 |
| FPT2W | | | 56 | | | | | 48 |
| FPT3E | | | 63 | | | | | 48 |
| FPT3N | | | 68 | | | | | 57 |
| FPT3W | | | 58 | | | | | 38 |
| CUTTC | | 63 | | | | | | 47 |
| CUTCD | | 57 | | | | | | 47 |
| CUTFL | | 49 | | | | | | 45 |
| CUTOL | | 41 | | | | | | 30 |
| CUTHO | | 49 | | | | | | 35 |
| MPT-A | 44 | 48 | | | | | | 54 |
| MPT-B | 40 | | | | | | | 43 |
| MPT-C | 43 | 45 | | | | | | 50 |
| SRT-A | | | | | | 93 | | 100 |
| SRT-B | | | | | | 61 | | 54 |
| FRT-A | | | | | | | 75 | 75 |
| FRT-B | | | | | | | 62 | 52 |
| RMTID | 59 | | | | | | | 55 |
| RMTDI | 65 | | | | | | | 55 |
| FOTIN | | | | | 90 | | | 100 |
| FOTRO | | | | | 55 | | | 60 |
| FFTAR | | | | 66 | | | | 72 |
| FFTHO | | | | 73 | | | | 72 |
| OFTWH | 54 | | | | | | | 45 |
| OFTBL | 54 | | | | | | | 36 |
| IPT | 69 | | | | | | | 57 |
| CRT | | | | | 42 | | | 48 |
| HPT | 50 | | | 45 | | | | 61 |
| GCT | | | | | | | | 14 |
| GT | | 54 | | | | | | 46 |
| % VAR | 11.1 | 10.6 | 10.4 | 6.0 | 5.4 | 5.0 | 4.1 | |

Note. Decimals and values less than .40 have been omitted. PSPD = perceptual speed; INOR = information ordering; CGPR = cognitive-perceptual ; FCLO = flexibility of closure; SPOR = spatial orientation; DEDR = deductive reasoning; INDR = inductive reasoning; $h^2$ = communality; % VAR = percentage of variance accounted for (total = 52.6).

not load highly (.32) on this factor. All the remaining variables load less than .30 on the factor.

The remaining four factors are defined solely by the test(s) and test sections of the constructs they were designed to measure. Factor four comprises the FFT and its model, the HPT, and is interpreted as a flexibility of closure (FCLO) ability. Factor five is defined by the FOT and its model, the CRT, and is interpreted as a spatial orientation (SPOR) ability. The sixth and seventh factors are defined by the SRT and FRT sections and are interpreted to be deductive reasoning (DEDR) and inductive reasoning (INDR) abilities, respectively.

The MPT sections load on the PSPD factor and two sections load on the INOR factor. All three sections of the MPT are highly speeded (see Table 8). The MPT is also similar to the less difficult sections of the CUT. Both tests require the application of a limited set of straightforward rules and the utilization of a schematic figure to produce correct responses. Only FPT1S and the GCT do not load on any factor. The highest loadings for the FTP1S is .37 on the INOR factor and .32 on the CGPR factor. The highest loading for the GCT is .23 on the FCLO factor.

Table 10 presents the results of the second factor analysis based on the order of presentation of the test sections. The results are very similar to the identity factor analysis, yielding seven, interpretable factors that account for 50.5 percent of the variance. There are some differences in the factors that are obtained and in the relative proportion of variance accounted for by the factors. First, the CGPR ability factor accounts for the most variance in the order analysis. Second, the INOR ability factor does not appear in the order analysis even though the CUT sections are presented in a fixed order (i.e., the order scores are equal to the identity scores). In contrast to the other tests that show a

Table 10

Factor Analysis of the NCORR Scores Based on Section Order of Presentation

| Section | CGPR | PSPD | FCLO | SPOR | PSEN | DEDR | INDR | $h^2$ |
|---|---|---|---|---|---|---|---|---|
| FPT-1 | | | | | | | | 37 |
| FPT-2 | 64 | | | | | | | 48 |
| FPT-3 | 60 | | | | | | | 45 |
| FPT-4 | 68 | | | | | | | 61 |
| FPT-5 | 56 | | | | | | | 39 |
| FPT-6 | 63 | | | | | | | 44 |
| FPT-7 | 65 | | | | | | | 49 |
| CUT-1 | | | | | | | | 40 |
| CUT-2 | | | 45 | | | | | 47 |
| CUT-3 | | | | | | | | 43 |
| CUT-4 | | | | | | | | 30 |
| CUT-5 | | | | | | | | 28 |
| MPT-1 | | | | | 47 | | | 56 |
| MPT-2 | | | | | 56 | | | 57 |
| MPT-3 | | 40 | | | 59 | | | 64 |
| SRT-1 | | | | | | 70 | | 64 |
| SRT-2 | | | | | | 65 | | 54 |
| FRT-1 | | | | | | | 90 | 93 |
| FRT-2 | | | | | | | 50 | 46 |
| RMT-1 | | 47 | | | | | | 37 |
| RMT-2 | | 62 | | | | | | 55 |
| FOT-1 | | | | 62 | | | | 61 |
| FOT-2 | | | | 85 | | | | 93 |
| FFT-1 | | | 42 | | | | | 44 |
| FFT-2 | | 42 | 42 | | | | | 46 |
| OFT-1 | | 54 | | | | | | 44 |
| OFT-2 | | 59 | | | | | | 44 |
| IPT | | 71 | | | | | | 61 |
| CRT | | | | 49 | | | | 53 |
| HPT | | 53 | 50 | | | | | 64 |
| GCT | | | | | | | | 16 |
| GT | | | 56 | | | | | 54 |
| % VAR | 12.2 | 10.8 | 6.6 | 6.0 | 5.7 | 4.8 | 4.3 | |

Note. Decimals and values less than .40 have been omitted. CGPR = cognitive-perceptual; PSPD = perceptual speed; FCLO = flexibility of closure; SPOR = spatial orientation; PSEN = problem sensitivity; DEDR = deductive reasoning; INDR = inductive reasoning; $h^2$ = communality; % VAR = percentage of variance accounted for (total = 50.5).

consistent improvement across sections, the CUT shows a general decrement across sections. The INOR ability factor is replaced by a factor that is defined by the three sections of the MPT and is interpreted to be a problem sensitivity (PSEN) ability. It is again recognized that other tests that are similar to the MPT have been designed to measure other ability constructs (e.g., perceptual foresight, spatial scanning; Hills, Douglas, and Lassiter, 1983).

Third, there are more sections in the order analysis that do not load on any factor or that load on two factors. Four of the five CUT sections do not load on any of the factors. As noted above, this may reflect the lack of a cumulative practice effect across sections. As in the identity factor analysis, the GCT and the first section of the FPT do not load on any factor. The third section of the MPT loads on both PSEN and PSPD. The HPT and the second section of the FFT load on both FCLO and PSPD, indicating the importance of speed as well as flexibility of closure on these tests. The loading of only the last sections of the MPT and FFT on the PSPD factor may indicate that performance on these tests reaches asymptotic levels within three and two practice iterations, respectively.

Finally, the third factor is somewhat difficult to interpret. It is labeled as FCLO since both the HPT and FFT sections have substantial loadings on the factor. However, the ASVAB GT score has the highest single loading on the factor. The relationship between the measures of flexibility of closure and the GT is not clear.

Test factor analyses. The identity and order factor analyses indicate that the sections of each test generally share a common, underlying factor. The following three factor analyses were conducted to determine the factor structure of the NCORR, %ACC, and CFG measures that were computed for each test. The analysis of the NCORR measure resulted in a three-factor solution that accounts for 51.7 percent of the variance (see Table 11). The first factor is interpreted to be a general, cognitive-perceptual ability factor that is primarily defined by the CUT, GT, FPT, and MPT. The two reasoning measures, the two flexibility of closure measures, and one spatial orientation measure also load on the CGPR factor.

The second factor is interpreted to be a perceptual speed ability factor, defined primarily by the IPT, RMT, and the OFT. The MPT, FFT, and HPT have substantial loadings on the PSPD factor as well as on the CGPR factor. The third factor is defined by the CRT and FOT, and is interpreted to be a distinct spatial orientation ability factor. The FOT also loads on the CGPR factor. The GCT is the only test that does not load greater than .19 on any factor.

The second factor analysis of test measures was based on the CFG scores. It is readily apparent (see Table 12) that correcting the NCORR scores for the effects of differential guessing does not affect the underlying factor structure. The analysis resulted in a three-factor solution that accounted for 52.5 percent of the variance. The factor loadings are virtually identical to the NCORR analysis. The three factors are interpreted as cognitive-perceptual, perceptual speed, and spatial orientation abilities, respectively.

The factor analysis of the %ACC scores resulted in a substantially different, three-factor solution that accounted for 42.0 percent of the variance (see Table 13). The first factor is interpreted to be a power factor (i.e., tests with relatively low accuracy; see table 6) that is defined by the CUT and FPT. The GT also loads highly on this factor, but no information is available on the accuracy rate of the GT that would be useful in interpreting the factor. The second factor is interpreted as a speed factor (i.e., tests with relatively high accuracy) that is defined by the spatial orientation and flexibility of closure tests. This interpretation is based on the nature of the scores (percentage of correctly

Table 11

Factor Analysis of the NCORR Scores for Each Test

| Test | CGPR | PSPD | SPOR | $h^2$ |
|---|---|---|---|---|
| FPT | 58 | | | 50 |
| CUT | 74 | | | 65 |
| MPT | 53 | 47 | | 58 |
| SRT | 45 | | | 30 |
| FRT | 42 | | | 33 |
| RMT | | 70 | | 62 |
| FOT | 43 | | 51 | 55 |
| FFT | 47 | 47 | | 49 |
| OFT | | 67 | | 52 |
| IPT | | 71 | | 60 |
| CRT | | | 94 | 100 |
| HPT | 40 | 53 | | 53 |
| GCT | | | | 08 |
| GT | 67 | | | 48 |
| % VAR | 20.0 | 19.7 | 12.0 | |

Note. Decimals and values less than .40 have been omitted. CGPR = cognitive-perceptual factor; PSPD = perceptual speed; SPOR = spatial orientation; $h^2$ = communality; % VAR = percentage of variance accounted for (total = 51.7%).

Table 12

Factor Analysis of the CFG Scores for Each Test

| Test | CGPR | PSPD | SPOR | $h^2$ |
|---|---|---|---|---|
| FPT | 64 | | | 55 |
| CUT | 75 | | | 66 |
| MPT | 54 | 44 | | 57 |
| SRT | 48 | | | 31 |
| FRT | 48 | | | 38 |
| RMT | | 64 | | 56 |
| FOT | 45 | | 53 | 57 |
| FFT | 48 | 47 | | 50 |
| OFT | | 64 | | 49 |
| IPT | | 76 | | 63 |
| CRT | | | 93 | 100 |
| HPT | 42 | 53 | | 55 |
| GCT | | | | 08 |
| GT | 67 | | | 49 |
| % VAR | 21.9 | 18.6 | 12.0 | |

Note. Decimals and values less than .40 have been omitted. CGPR = cognitive-perceptual factor; PSPD = perceptual speed; SPOR = spatial orientation; $h^2$ = communality; % VAR = percentage of variance accounted for (total = 52.5%).

Table 13

Factor Analysis of the %ACC Scores for Each Test

| Test | POWER | SPEED | DEGR | $h^2$ |
|---|---|---|---|---|
| FPT | 67 | | | 56 |
| CUT | 72 | | | 63 |
| MPT | 49 | | | 37 |
| SRT | | | | 34 |
| FRT | 49 | | | 37 |
| RMT | | | | 22 |
| FOT | | 66 | | 58 |
| FFT | | 44 | | 34 |
| OFT | | | 71 | 60 |
| IPT | | | | 21 |
| CRT | | 71 | | 60 |
| HPT | | 50 | | 38 |
| GCT | | | 45 | 23 |
| GT | 63 | | | 47 |
| % VAR | 17.5 | 14.4 | 10.1 | |

Note. Decimals and values less than .40 have been omitted. POWER = low accuracy factor; SPEED = high accuracy factor; DEGR = accuracy using degraded stimuli; $h^2$ = communality; % VAR = percentage of variance accounted for (total = 42.0%).

attempted items) rather than the test loadings, per se. The RMT and IPT, both measures of perceptual speed, do not load on any of the factors.

The third factor is difficult to interpret. Although the OFT and GCT were designed to measure speed of closure, the previous results indicate the OFT is more similar to the RMT (i.e., a measure of perceptual speed) than it is to the GCT. The third factor is not interpreted as speed of closure since the OFT and GCT have not shown a substantial relationship in any of the previous analyses. The factor is also not interpreted as representing another level of accuracy. The mean %ACC scores for the OFT and GCT are within the range of scores associated with the high accuracy tests (see Table 8).

Two other tests load on the factor at near the .40 criterion. Unfortunately, they do not readily clarify the interpretation. The IPT (.39 loading) has the highest mean %ACC, which would support a speed interpretation. However, the stimulus figures used on the IPT are small and of medium reproduction quality, which could support a speed of closure interpretation. The SRT (.37 loading) has a relatively low %ACC (i.e., the SRT is a power test; it also loaded .37 on factor 1) and is verbal in content rather than figural. None of the other tests have a loading greater than .27.

The only communality among the four tests appears to be the degraded condition of the stimulus materials. This is a specified condition with the OFT and GCT. The relatively few incorrect responses on the IPT are consistently made on the items which have the lowest graphics quality. Finally, the major and minor terms in the SRT syllogisms are nonsense syllables. The lack of any inherent meaning in the terms and in the syllogisms could be interpreted as degraded information. Therefore, the third factor is interpreted as accuracy of performance when the stimuli are degraded (DEGR).

Regression and gain score factor analysis. The final factor analysis was based on the NCORR identity scores for the experimental tests and the two regression and four gain score measures derived from the FPT. The initial analysis resulted in a singular matrix because of the correlation between the FPT NCORR and the two regression measures. The intercept of the regression equation was deleted from the analysis. The subsequent analysis resulted in a five-factor solution that accounts for 52.3 percent of the variance (see Table 14). The first factor is defined by the RMT, IPT, and OFT and is interpreted as a perceptual speed ability. The second factor, defined by the CUT, GT, and FPT, is interpreted to represent a general, cognitive-perceptual ability. Both factors are consistent with the preceding results.

The third and fourth factors are defined by the level 3 and level 2 gain scores, respectively. The gain scores were designed to assess the effects of review and practice (i.e., learning) at two levels of difficulty. Although the gain scores within a level of difficulty are positively correlated (level 2 = .47 and level 3 = .57), the correlations between levels are essentially zero (none greater than -.07). The low correlations do not appear to be a function of the gain score distributions. All four distributions are approximately symmetrical and have substantial and homogenous variances. The gain scores at the two levels of difficulty apparently are not measuring the same underlying ability.

These two factors are, consequently, difficult to interpret. None of the other measures load greater than .13 on either factor. The only other information is the descriptive statistics on the four variables. Since there was a positive learning curve at the second level of difficulty, the fourth factor is interpreted as a learning (LRNG) ability factor. Since there was a slight decrease in performance over sections at the third level of difficulty, the third factor is tentatively interpreted as a fatigue (FATG) factor. Although there are other

Table 14

Factor Analysis of the NCORR, Regression, and Gain Scores for Each Test

| Score | PSPD | CGPR | FATG | LRNG | SPOR | $h^2$ |
|---|---|---|---|---|---|---|
| FPT | 44 | 53 | | | | 50 |
| CUT | | 72 | | | | 65 |
| MPT | 53 | 50 | | | | 59 |
| SRT | | 46 | | | | 31 |
| FRT | 41 | | | | | 33 |
| RMT | 74 | | | | | 63 |
| FOT | | 45 | | | 45 | 55 |
| FFT | 51 | 46 | | | | 50 |
| OFT | 69 | | | | | 52 |
| IPT | 73 | | | | | 59 |
| CRT | | | | | 90 | 100 |
| HPT | 57 | | | | | 53 |
| GCT | | | | | | 10 |
| GT | | 69 | | | | 52 |
| SLP[a] | | | | | | 07 |
| GS1 | | | | 99 | | 100 |
| GS2 | | | | 47 | | 22 |
| GS3 | | | 57 | | | 33 |
| GS4 | | | 100 | | | 100 |
| % VAR | 17.3 | 14.3 | 7.3 | 6.7 | 6.7 | |

Note. Decimals and values less than .40 have been omitted. PSPD = perceptual speed; CGPR = cognitive-perceptual factor; FATG = fatigue factor; LRNG = learning factor; SPOR = spatial orientation; $h^2$ = communality; % VAR = percentage of variance accounted for (total = 52.3%).

[a] Inclusion of the intercept resulted in a singular matrix; the SLP and INT are correlated -.81 and the INT and FPT are correlated .58.

reasonable interpretations (e.g., a cognitive overload factor), a differential fatigue interpretation would explain the low correlation between the gain scores and would be consistent with observations by the subjects. That is, subject critiques during the preliminary test phase and comments following the experimental administrations indicated that many subjects were exhausted by the first four or five sections of the FPT.

The fifth factor is defined solely by the CRT and FOT and is interpreted to represent a spatial orientation ability. The PSPD, CGPR, and SPOR factors are basically equivalent to the factors on the NCORR and CFG analyses. As in the other analyses, the GCT does not load greater than .22 on any of the obtained factors. Finally, the SLP does not load greater than .14 on any of the factors.

Although the SLP distribution is indicative of substantial individual differences, the scores do not appear to be reliable. The individual correlations between the level of difficulty and performance on each section range from -.98 to +.80. The standard errors are also relatively high, ranging from .52 to 3.69. It is apparent that the changes in performance within difficulty levels is confounding the assessment of changes in performance across difficulty levels. An attempt to rectify this confound by regressing the level of difficulty on the mean score for each level was unsuccessful. The resulting regression coefficients were too affected by performance on the first level of difficulty (i.e., performance on FPT1S is highly correlated with the slope (-.70) and the intercept (.96) of the reqression equations). The two regression measures do not appear to contribute any unique or reliable information to the evaluation of subject abilities.

## Summary and Recommendations

This research was conducted to evaluate a complex, cognitive-perceptual aptitude test and a battery of eight psychometric tests as potential predictors of performance in initial helicopter flight training. The psychometric characteristics of most of the tests are sufficient to justify further research on their utility as selection instruments. The average difficulty of the tests is near the optimum level of .50 while the item difficulties exhibit a range that is sufficient to assess all levels of ability. The test variances indicate there are substantial individual differences in subject performance. The estimates of reliability and item discriminability are also generally satisfactory.

Despite the overlap in method variance (i.e., all the tests are multiple choice, paper-and-pencil tests), the battery appears to be tapping seven independent dimensions of human abilities. When the test sections are combined into total scores, the battery assesses three primary ability domains: cognitive-perceptual, perceptual speed, and spatial orientation. Whether a representative test from each domain is sufficient for predicting helicopter pilot training will have to be determined in a validation study. Perhaps most importantly, there is evidence that scores may be derived from the Flight Planning Test (FPT) that are indices of learning ability. The learning indices are independent of the three primary ability factors obtained in the NCORR analysis. Each of the tests, as well as the sample of subjects on which the evaluations are based, are discussed below in more detail.

### Sample Characteristics

The evaluation of the FPT and the experimental battery is dependent upon the characteristics of the subjects who participated in the research. The subjects in this study are not highly representative of the current population of Army helicopter pilot students, although there are some similarities (e.g., performance on the Educational Testing Service (ETS) tests, age, military status). The primary areas of dissimilarity are performance on the General Technical (GT) test and educational level. Subjects that scored higher than 110 on the GT (approximately half of the sample) are generally representative of current Aviation Warrant Officer applicants. The majority of applicants (i.e., those who score at least 110 on the GT) meet the minimum test requirements for selection into helicopter pilot training.

Recognizing that the current project was not conducted to evaluate the predictive validity of the FPT or the battery, it is more important that the research sample be representative of the general population from which student pilots are drawn (i.e., U. S. Army recruits, junior enlisted personnel, and junior commissioned officers). The present sample meets this criterion on virtually all points of comparison, except for interest in the aviation field. Nearly one third of the subjects indicated they had no interest in aviation, but this may reflect an awareness by many subjects that they do not meet the very stringent physical standards that are currently required of aviation applicants.

Despite the diversity of interest in the aviation field, the subjects had reasonably high expectations of their ability to perform well on the tests and had a positive attitude toward taking the battery. The positive responses to the two attitudinal questions are particularly important in the evaluation of the battery. The tests in the experimental battery are relatively difficult and lengthy. Subjects who had little confidence in their abilities or who were not interested in taking the battery would no doubt contribute data of suspect value. Every reasonable effort was made to motivate the subjects to perform to the best of their abilities and, subsequently, to evaluate the data collected from subjects who failed to follow test procedures or did not give a reasonable effort .

In summary, performance on the experimental battery by this sample of research subjects is considered to be representative of the broad spectrum of abilities that exist among military personnel that are in the training stage of their careers. Furthermore, the subjects are considered to have been as highly motivated to demonstrate maximum performance as could be expected under the conditions of administration.

Evaluation of the Battery

One of the purposes of this research was to identify tests that could potentially be used to select helicopter pilot trainees. A study would then be conducted to validate performance on the tests against performance during flight training. Six of the eight psychometric tests in the experimental battery appear to be good candidates for inclusion in the validation battery. The Obscured Figures Test (OFT) and the Map Planning Test (MPT) are not recommended for further study. The ETS tests were included for research purposes only; they are not being considered for use as selection tests.

The characteristics of the OFT are the least acceptable of any test in the battery. Although the difficulty level, variability, and internal consistency of the OFT are within acceptable limits, the correlation between the two sections is very low (.38). The only difference between the two sections is the figure-ground reversal. More importantly, the OFT appears to be measuring the same construct (perceptual speed) as the Rapid Match Test (RMT) rather than assessing speed of closure. Since the other characteristics of the RMT are equivalent to or better than the OFT, it is recommended that the OFT be deleted from the validation battery.

Speed of closure ability has many applications for a pilot, especially for a military pilot. In navigating an aircraft, for example, the pilot may be required to identify unfamiliar landmarks that are partially obscured by vegetation or weather conditions (e.g., by fog or rain). The military pilot may also have to recognize unknown enemy positions and weapons that are intentionally camouflaged; in such instances, an insufficient speed of closure ability may be lethal.

Nonetheless, speed of closure ability is very difficult to measure using a multiple choice, paper-and-pencil format. Having multiple choice objects for comparison with the test stimulus changes the ability requirements for performance on the test. The standardized paper-and-pencil tests of the ability, such as the Gestalt Completion Test (GCT), employ an open-ended, verbal response format. Although this format eliminates the target-to-choice comparison problem, the open-ended format requires the ability not only to form a complete mental representation of the obscured object, but also the ability to name it. Prior experience or familiarity with the object and differences in vocabulary obviously influence performance on the test.

In addition, the open-ended format requires interpretation in scoring the test that may create another source of error variance. It should be noted, however, that scoring errors or the prior experiences of this sample did not differentially affect the GCT scores. Performance on the GCT by the present sample is very similar to the test performance of a large sample of flight students in a previous project. Despite the importance of speed of closure as a requirement for effective performance as a pilot, a multiple choice, paper-and-pencil test is not available to assess this ability.

The MPT is also a marginal contributor to the assessment of individual differences in the battery. Although the difficulty, variability, and reliability of the test are very satisfactory, it does not make a consistent and unique contribution to the factor stucture of the battery. Sections of the test load on both the perceptual speed (PSPD) and information ordering (INOR) factors in the identity analysis and on the problem sensitivity (PSEN) and PSPD factors in the order analysis (see Tables 9 and 10). The test loads on both the cognitive-perceptual (CGPR) and PSPD factors in the NCORR, CFG, and regression and gain score analyses (see Tables 11, 12, and 14). Although the MPT %ACC mean (88.23) indicates the test is highly speeded, it loads on the POWER factor in the %ACC analysis (see Table 13). Therefore, it is recommended that the MPT be deleted from the validation battery since the test appears to be primarily another measure of perceptual speed.

Although some modifications are needed, the remaining six tests in the experimental battery are recommended for further study in the validation battery. The Chart Use Test (CUT) requires the most modifications. It is the most difficult test (average percent correct = .39) in the battery and it has the lowest intercorrelations among the sections of the test (average $r$ = .40). As has been noted previously, the CUT sections are presented in an order of increasing difficulty. That is, the difficulty level of the total test is not representative of the easier sections. In addition, the sections are not designed to be equivalent. As a result, the difficulty level and between-section reliabilities are not considered to be unacceptable. In contrast, the CUT scores have the highest relative variability of any test in the battery, indicating that substantial individual differences are being assessed.

The CUT is consistently identified with the cognitive-perceptual factor except in the section analyses. The CUT defines a factor tentatively interpreted as information ordering in the identity analysis, but it does not load on any factor in the order analysis. Overall the CUT seems to be most closely related to the GT composite from the Armed Services Vocational Aptitude Battery (ASVAB). That is, the CUT is essentially a general mental ability test that requires reading facility, paragraph comprehension, and mathematical ability. General cognitive ability is important for virtually all training and learning situations (Hunter & Hunter, 1984), but especially for such a complex program as helicopter pilot training. In addition, the CUT is, to some degree, content valid since the graphs and procedures are modeled on graphs in the utility helicopter operating manual. It is recommended that the CUT be revised by allowing more time on each section and by deleting the most difficult section of the test. The revised CUT should be studied further to determine whether the test enhances the prediction of pilot training performance or is redundant of the GT.

The two measures of reasoning ability exhibit very acceptable psychometric characteristics, although the Sound Reasoning Test (SRT) has the highest percentage (.16) of negative item-total correlations. None of the negatively correlated items are miskeyed. The negative correlations may be the result of guessing on the more difficult items by the less capable subjects. The SRT and the Finding Rules Test (FRT) each define a unique factor in the section analyses, even though recent researchers (e.g., Colberg, Nester, & Trattner, 1985) have argued for a convergence of the inductive and deductive reasoning abilities. Different abilities are obviously required for performance on the SRT and FRT.

Both deductive and inductive reasoning ability are required during initial training and in flying the aircraft. Jensen (1982), among others, has noted that the decision-making workload of modern pilots is increasing while the psychomotor workload is decreasing. Therefore, both tests are recommended for further study in the validation battery. However, the time limits for each test should be slightly more restrictive to reduce the total administration time and, consequently, to reduce fatigue. The tests do not appear to be affected by the current time limits and moderately restrictive time limits do not appreciably affect the psychometric characteristics of power tests (Nunnally, 1978, p. 638).

It may also be possible to derive gain scores from these tests as indices of learning. The average score on the second section of each test showed a significant improvement over the first section. Although the other tests in the battery also showed significant practice effects, only the SRT and FRT are designed to have equivalent sections that could be used to derive gain scores. Gain scores derived from the other tests in the battery would be confounded by the inequality of difficulty levels unless the sections were presented in a fixed order of ascending difficulty.

The remaining three tests exhibit excellent psychometric characteristics, although it is recognized that they are highly speeded tests and that the estimates of internal consistency are somewhat inflated. The correlations between the sections are very high even though the tests are designed to have some differences between the sections. The construct validity of the tests is supported by the correlations between the tests and their Educational Testing Service (ETS) models. In each case, the test is more highly correlated with its ETS model than with any other test.

The RMT and Figure Orientation Test (FOT) consistently define unique factors in the analyses of the battery. This is consistent with the literature on the structure of human abilities; perceptual speed and spatial orientation are two of the most strongly supported abilities in the various taxonomies (e.g., Dunnette, 1976; Fleishman & Quaintance, 1984). Both abilities are requirements for effective performance as pilots. Pilots of modern aircraft constantly have to scan their instruments and the visual scene outside the aircraft. Both speed and accuracy are critical in this rapidly evolving environment. Spatial orientation is also extremely important in navigation and in the control of aircraft that are capable of movement about three rotational axes. Both tests are recommended for further study in the validation battery. However, more restrictive time limits should be used with the RMT to increase the mean difficulty level and reduce the total administration time.

The Finding Figures Test (FFT) defines a unique factor only in the analyses of the test sections. In the other factor analyses, the FFT has approximately equal loadings on both PSPD and CGPR, indicating that the test has both a speeded component and a cognitive component. The cognitive component may be related to the substantial differences in difficulty of the two FFT sections. Flexibility of closure is again an important ability for piloting an aircraft. Known landmarks, topographical features, and targets are frequently embedded in vegetation or obscured by weather or light conditions. The FFT is also recommended for further study as part of the validation battery.

It is also important that the three tests are only moderately intercorrelated (see Table 5), despite the similarities in their construction. For example, each of the tests is highly speeded and uses similar stimulus materials, especially the figures in the RMT and FOT. Nonetheless, each of the tests appears to be assessing unique sources of variance in subject abilities. Further research is needed to determine whether the tests will make unique contributions to the prediction of performance in initial helicopter flight training.

Evaluation of the CFG and %ACC measures

The mean corrected-for-guessing (CFG) and percent accuracy (%ACC) scores are both computed from the number correct (NCORR) scores and the mean number of items attempted. As such, the descriptive statistics are of limited value, but the interrelationships among the individual scores may reflect changes in the factor structure of the battery as a function of differential guessing or speed-versus-accuracy tradeoffs. However, the factor analysis of the CFG scores is virtually identical to the NCORR analysis. This result is consistent with the literature (e.g., Nunnally, 1978, pp.644-650) on corrections for guessing when subjects are instructed to respond to all items or to not guess at all. The subjects in this study were instructed to attempt every item on which they held any reasonable opinion. The subjects were discouraged from making random guesses, but they were told there would be no penalty for erroneous responses. As a result, differential guessing did not affect test performance.

The %ACC factor analysis yields a simpler structure (i.e., fewer variables load on two factors) than the NCORR, but the solution accounts for less variance. The %ACC analysis clearly distinguishes between the power tests and the moderately speeded tests in the battery, although the loadings on these factors are very similar to the CGPR and PSPD factors in the NCORR and CFG analyses. In addition, a spatial factor is not obtained in the %ACC analysis; the spatial orientation measures load on the speed factor. The third factor is, with considerable difficulty, interpreted as accuracy of performance under degraded conditions. Although this factor appears to be a unique perspective on the individual differences under investigation, there is insufficient evidence to support the interpretation. The calculation of %ACC scores does not appear to assess a reliable and unique source of variance in the analysis of individual ability differences.

Evaluation of the FPT

The evaluation of the FPT indicates that the test generally fulfills the design requirements for an aviation-related, complex cognitive-perceptual test. The FPT should definitely be retained for further study in the validation battery. Before discussing the statistical results, there are several observations of subject reactions and comments from the subject critiques during the preliminary test phase that are relevant to the design requirements. First, the subjects believed that the test is highly related to helicopter flight training. That is, the test appears to have "face" validity. This perception was generally held by nonaviation subjects, current flight student subjects, and experienced pilots. In contrast, nearly all subjects reacted negatively to the FRT, which is quite abstract and has no apparent relationship to aviation.

Second, there is no direct evidence that any of the subjects were so unfamiliar with the terms or concepts (e.g., map coordinates, flight restrictions) used in the test that the lack of prior experience was detrimental to their performance. The subjects were provided ample opportunities to ask questions, but none ever asked for clarification of the test terms or concepts. The instructions, practice questions, and performance feedback also appear to be adequate to enable the subjects to work the airspeed and flight time questions, even though there are probably large differences in the mathematical backgrounds of the subjects. Concerns expressed during the peer review that the FPT was too difficult for administration in the general recruit population appear to have been overcome by improved instructions, examples, and practice.

Finally, the FPT is suitable for administration to large groups. There were no problems encountered in administering the FPT that are attributable to differences in group size (sample sizes ranged from 9 to 155). The test is too lengthy, however. Although the test time is 63 minutes, the total administration time is approximately 90 minutes. Many subjects became fatigued after the fifth or sixth section of the FPT was administered.

Psychometric evaluation. The FPT is slightly more difficult (mean difficulty level = .41) than the optimal level of .50, but two factors mitigate this finding. First, performance may be expected to improve in a more restricted sample. That is, in operational use the FPT will be administered only to applicants who have, at a minimum, completed high school and scored 110 or greater on the ASVAB GT. As shown in Table 4, the GT score is positively correlated with performance on the FPT. Second, the FPT is designed to have an increasing level of difficulty across sections. The mean difficulty level is .52 for the first section of the FPT. The difficulty level increases to an average of .35 for the last three sections. As previously noted, fatigue may have affected performance on the later sections of the FPT as well. In general, the difficulty level of the FPT is considered to be acceptable, if not optimal.

The variability of the FPT scores is indicative of substantial individual differences in the underlying abilities. The variability of the scores based on the section identity and on the section order of presentation is both high and homogeneous. The standard deviations of the sections are not significantly altered by changes in the level of difficulty or the amount of practice. Although the internal consistency of the FPT is high, the correlations between the sections are moderate. This is interpreted to mean that a unique source of variance is being assessed by each of the sections, at least in a pairwise comparison. That is, although the intersectional correlations are high enough to indicate the sections are tapping a common construct, they are low enough to indicate that the sections are not completely redundant.

The pattern of differences among the identity section means is indicative of the successful implementation of the design specifications. With one exception, performance on the sections is significantly different between levels of difficulty but not significantly different within levels of difficulty. The same pattern generally holds for the order-of-presentation section means between levels of difficulty. Within the levels of difficulty, however, only performance on difficulty level 2 is consistent with the design specifications. After an initial drop in performance from the first level of difficulty, performance steadily increases over the subsequent two sections. Performance on difficulty level 3 shows an initial drop from level 2 performance, but there is no improvement as a result of additional study and practice. The lack of a practice effect at difficulty level 3 is attributed to subject fatigue rather than the excessive difficulty of the sections.

Each of the six item types contributes to the total score on the FPT, although the different types assess different combinations of abilities. Items of each type correlate greater than .30 with the total score a minimum of three and a maximum of eight times. Mathematical ability does not appear to be a major requirement on the FPT, although it is certainly one of the complex of abilities. Although the mathematical computations are equivalent, the flight time items correlate .30 with the total score the maximum number of times while the airspeed items correlate .30 with the total score the minimum number of times. These two item types are the most frequently skipped, but the number of skipped items is not excessive. Many more subjects failed to attempt the last four items than skipped the four airspeed and flight time items on each section.

The ability to disregard irrelevant information does appear to be an important factor in performance on the FPT. Nearly half of the items that contained "selective attention" information were highly correlated (greater than .30) with the total score. Although irrelevant or erroneous information is essential in multiple choice alternatives, the inclusion of irrelevant information in the stem is contrary to most principles of item construction. The contribution of these items to the total scores should be closely scrutinized during the validation study. Most of the highly correlated, selective attention items are associated with the compass-heading and number-of-turns item types. Both of these item types are designed to require spatial ability. Further research is also needed on the, admittedly speculative, interaction between selective attention ability and spatial ability.

Factor analytic evaluation. In both the identity and order-of-presentation analyses, six of the seven sections of the FPT load on a single factor that is interpreted as a complex, cognitive-perceptual ability. Section 1S, representing the lowest level of difficulty, does not load .40 or greater on any factor. The loadings of only the more complex FPT sections support the complexity of the underlying ability construct. Section 1S is much simpler than any other section in terms of the number of item types, the presence of selective attention information, and the complexity of the route map. Furthermore, the multiple abilities that are represented by the factor are evident in the NCORR factor analysis. In that analysis, nearly all of the variables except the spatial orientation and perceptual speed measures load on the CGPR. The multiple test loadings are considered to be more important than the absolute magnitude of the CUT and GT loadings. The factor loadings of the CUT and GT are less consistent than the FPT loadings across the analyses.

There are, of course, other interpretations of the CGPR factor. Since most of the tests have a high loading on the factor, it could be interpreted as a general intelligence factor (cf. Spearman, 1904). This interpretation of the factor is supported by the loading of the GT score, which has frequently been interpreted as a general mental ability test. However, general mental ability or general intelligence is structurally analagous to cognitive-perceptual ability; i.e., each label represents a complex of individual human abilties. The important distinction is that the standardized mental ability tests, such as the GT, depend upon the two critical assumptions of equal prior experience and the representativeness of a static measure as an index of a dynamic process. The FPT measures a complex of cognitive and perceptual abilities, permits the demonstration of learning, and minimizes the effects of prior experience. The cognitive-perceptual label for the factor is intended to represent this distinction.

The results of the factor analysis of the intercept (INT) and slope (SLP) of the regression equations and of the four gain scores (GS1-GS4) are mixed.

The INT was deleted from the analysis because the combination of INT, SLP, and FPT scores produced a singular matrix. In the subsequent analysis, the SLP does not load on any factor. As was previously noted, the SLP scores do not appear to be reliable. The correlations between section scores and levels of difficulty range from high positive to high negative. If the scores were reliable, the correlations should range only from low to high negative. The standard errors are also unacceptably high for a majority of the regression equations. The decrement in performance across levels of difficulty is apparently confounded by the effects of practice within the second level of difficulty. In conclusion, the INT is a redundant score and the SLP score appears to be an unreliable measure of the individual differences in ability.

The gain scores from the second level of difficulty appear to be reliable indices of learning as a function of study and practice. The variance of both scores is relatively high and homogeneous, and the distribution of the two scores is somewhat platykurtic. If there were few individual differences in learning, the expected distribution would have a mean of 12, low variance and a leptokurtic shape. The average for GS1 is higher than the expected value, indicating an improvement in performance with a single iteration of study and practice. GS2 is significantly higher than GS1, indicating a further learning effect with a second iteration of study and practice. The moderate (.47) correlation between the two gain scores indicates that additional variance can be accounted for by the second iteration. Nonetheless, the first two gain scores define a common factor in the NCORR, regression, and gain score analysis.

The gain scores from the third level of difficulty have approximately the same distributions as GS1 and GS2 except for the mean values. GS3 shows a slight decrement from the expected value and GS4 is approximately equal to GS3. GS3 and GS4 are more highly correlated with each other than the first two scores, indicating more common variance (33 percent versus 22 percent) in the level 3 scores. The central tendency of the latter two gain scores does not support their interpretation as indices of learning. Furthermore, the difficulty level 2 gain scores are essentially uncorrelated with the difficulty level 3 gain scores, and they load on different factors. The factor defined by GS3 and GS4 is interpreted to be a fatigue factor. This interpretation of the factor is partially based on the observation of subjects during the test administration and from the subject critiques collected during the preliminary phase of testing.

Most importantly, the level 2 and level 3 gain scores are independent of the FPT scores on which they are based. The FPT does not load on the LRNG or FATG factors, nor do the gain scores load on the CGPR factor. As such, the FPT total score and the gain scores represent unique measures of individual differences.

FPT recommendations. The FPT has been the focus of this research as a unique approach to assessing individual differences in a complex of abilities. The psychometric and factor analytic evaluations indicate that the FPT should be studied further to determine its utility for predicting the performance of helicopter pilot trainees. It is recommended, however, that the FPT be reduced from seven to six sections before further research is conducted. After an initial drop in performance from the fifth section, performance on the last two sections of the FPT is very similar in level and moderately correlated. That is, little additional change occurs as a result of the last study and practice iteration. In addition, the observed fatigue among the subjects necessitates a reduction in the length of the test.

Furthermore, it is recommended that section 3N be deleted from the test. The average score on section 3N is anomalous within and across the levels of difficulty. That is, the average score on section 3N is significantly different from the other level 3 sections but not different from two of the level 2 sections. The anomaly should not have significantly affected other analyses because the order of presentation of the sections was counterbalanced. The revised FPT should then be studied further as part of the validation battery.

In summary, the results of this study indicate that the FPT and six of the eight tests in the psychometric battery are suitable, with minor modifications, for further research as predictors of helicopter pilot training performance. The FPT appears to assess reliable individual differences in multiple cognitive and perceptual abilities under different levels of complexity. The subjects in the study believed the FPT to be relevant to flight training and gave no indication that prior experience differentially affected their performance. The FPT also appears to yield a static measure of ability (total score) and dynamic measures of learning and fatigue. Whether these measures are predictive of IERW performance will have to be ascertained in the validation study.

References

Allen, G. L., Secunda, M. D., Salas, E., & Morgan, B. B., Jr. (1982). Evaluation of rate parameters as predictors of the acquisition, decay, and reacquisition of complex cognitive skills (Technical Report No. ITR-82-27). Norfolk, VA: Old Dominion University, Center for Applied Psychological Studies.

Allen, G. L., & Morgan, B. B., Jr. (1984). Assessment of learning abilities using rate measures. Proceedings of the Ninth Psychology in the Department of Defense Symposium, 559-563.

American Psychological Association (1981). Ethical principles of psychologists (revised). American Psychologist, 36, 633-638.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. Personnel Psychology, 27, 519-533.

Barrett, G. V., Alexander, R. A., Cellar, D., Doverspike, D., & Thomas, J. C. (1983). Use of an information-processing based test battery in an applied setting: Prediction of monitoring performance. Perceptual and Motor Skills, 56, 939-945.

Bejar, I. I. (1985). Speculations on the future of test design. In Embretson, S. E. (Ed.), Test design: Developments in psychology and psychometrics. Orlando, FL: Academic Press.

Boehm, V. R. (1982). Are we validating more but publishing less? (The impact of governmental regulation on published validation research - an explanatory investigation). Personnel Psychology, 35, 175-187.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "Structure of Intellect." In L. B. Resnick (Ed.), The nature of intelligence. Hillsdale, NJ: Erlbaum.

Carroll, J. B. (1978). How shall we study individual differences in cognitive abilities? - - Methodological and theoretical perspectives (Technical Report

No. 1). Chapel Hill, NC: The L.L. Thurstone Psychometric Laboratory, University of North Carolina.

Cascio, W. F. & Phillips, N. F. (1979). Performance testing: A rose among thorns? Personnel Psychology, 32, 751-766.

Christal, R. E. (1976, October ). What is the value of aptitude tests? Paper presented at the Annual Conference of the Military Testing Association, Gulf Shores, AL.

Christal, R. E., Tirre, W. C., & Kyllonen, P. C. (1984). Two for the money - - speed and level scores from a computerized vocabulary test. Proceedings of the Ninth Psychology in the Department of Defense Symposium, 553-557.

Colberg, M., Nester, M. A., & Trattner, M. H. (1985). Convergence of the inductive and deductive models in the measurement of reasoning abilities. Journal of Applied Psychology, 70, 681-694.

Comrey, A. L. (1973). A first course in factor analysis. New York: Academic Press.

Cronbach, L. J. (1970). Essentials of psychological testing (3rd ed.). New York: Harper & Row.

Department of Defense (1985). Defense 85 almanac. Arlington, VA: Department of Defense.

Dixon, W. J. (Chief Ed.), Brown, M. B., Engelman, L., Frane, J. W., Hill, M. A., Jennrich, R. I., & Toporek, J. D. (1983). BMDP statistical software. Los Angeles: University of California Press.

Dunnette, M. D. (1976). Aptitudes, abilities, and skills. In Dunnette, M. D. (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Ekstrom, R. B., French, J. W., & Harmon, H. H. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Embretson, S. E. (Ed.), (1985). Test design: Developments in psychology and psychometrics. Orlando, FL: Academic Press.

Farell, B. (1985). "Same" - "different" judgments: A review of current controversies in perceptual comparisons. Psychological Bulletin, 98, 419-456.

Fleishman, E. A. (1975). Toward a taxonomy of human performance. American Psychologist, 30, 1127-1149.

Fleishman, E. A., & Quaintance, M. K. (1984). Taxonomies of human performance. Orlando, FL: Academic Press.

Freyd, M. (1923). Measurement in vocational selection: An outline of research procedure. Journal of Personnel Research, 2, 215-249, 268-284, 377-385.

Gettinger, M. & White, M. A. (1979). Which is the stronger correlate of school learning? Time to learn or measured intelligence. Journal of Educational Psychology, 71, 405-412.

Ghiselli, E. E. (1966). The validity of occupational aptitude tests. New York: Wiley.

Ghiselli, E. E. (1973). The validity of aptitiude tests in personnel selection. Personnel Psychology, 26, 461-477.

Guilford, J. P., & Hoepfner, R. (1971). The analysis of intelligence. New York: McGraw-Hill.

Guion, R. M. (1965). Personnel testing. New York: McGraw-Hill.

Guion, R. M. (1976). Recruiting, selection, and job placement. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Gutenberg, R. L., Arvery, R. D., Osburn, H. G., & Jeanneret, P.R. (1983). Moderating effects of decision-making/information-processing job dimensions on test validities. Journal of Applied Psychology, 68, 602-608.

Hills, J. R., Douglas, K., & Lassiter, K. (1983). Measurement of deductive reasoning, inductive reasoning, and sensitivity to problems for predicting success in helicopter pilot training (Third Interim Report). Tallahassee, FL: Florida State University.

Hunt, E. B. (1976). Varieties of cognitive power. In Resnick, R. B. (Ed.), The nature of intelligence. Hillsdale, NJ: Erlbaum.

Hunt, E. B. (1980). Intelligence as an information processing concept. British Journal of Psychology, 71, 449-474.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

Imhoff, D. L., & Levine, J. M. (1981, January). Perceptual-motor and cognitive performance task battery for pilot selection (AFHRL Technical Report 80-27). Brooks Air Force Base, San Antonio, TX: Air Force Human Resources Laboratory.

Jensen, R. S. (1982). Pilot judgment: Training and evaluation. Human Factors, 24, 61-73.

Kantor, J. E., & Bordelon, V. P. (1985). The USAF pilot selection and classification research program. Aviation, Space, and Environmental Medicine, 56, 258-261.

Maier, M. H., & Grafton, F. C. (1981). Aptitude composites for ASVAB 8, 9, and 10 (Research Report No. 1308). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McAnulty, D. M., Jones, D. H., Cohen, R. J., & Lockwood, R. E. (1984). Identification of the abilities required for effective helicopter training performance (Technical Report ASI479-046-84). Fort Rucker, AL: Anacapa Sciences, Inc.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984, October). Validation of current and alternative Armed Services Vocational Aptitude Battery (ASVAB) area composites (Technical Report 651). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Myers, D. C., Schemmer, F. M., & Fleishman, E. A. (1982). Analysis of computer interactive tests for assigning helicopter pilots to different missions (Technical Report R83-8). Washington, DC: Advanced Research Resources Organization.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Onken, J., Hastie, R., & Revelle, W. (1985). Individual differences in the use of simplification strategies in a complex decision-making task. Journal of Experimental Psychology: Human Perception and Performance, 11, 14-27.

Passey, G. E. & McLaurin, W. A. (1966). Perceptual-psychomotor tests in aircrew selection: Historical review and advanced concepts (PRL-TR-66-4). Lackland AFB, TX: Personnel Research Laboratory.

Payne, D. L., & Tirre, W. C. (1984). Individual differences in learning rate. Proceedings of the 9th Psychology in the Department of Defense Symposium, 548-552.

Rossmeissl, P. G., Martin, C. J., & Wing, H. (1983, October). Validity of ASVAB 8/9/10 as predictors of training success. Paper presented at the 25th Annual Conference of the Military Testing Association, Gulf Shores, AL.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. American Psychologist, 36, 1128-1137.

Schmidt, F. L., & Hunter, J. E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. Journal of Applied Psychology, 68, 404-414.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. (1979). The impact of valid selection procedures on work force productivity. Journal of Applied Psychology, 64, 609-626.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 32, 257-281.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on workforce productivity. Personnel Psychology, 35, 333-347.

Schmidt, F. L., Pearlman, K., Hunter, J. E., & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis. Personnel Psychology, 38, 697-799.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.

Smith, M. G., Krause, M., Kennedy, R. S., Bittner, A. C., & Harbeson, M.M. (1983). Performance testing with microprocessors: Mechanization is not implementation. Proceedings of the 27th Annual Meeting of the Human Factors Society, 1, 674-678.

Spearman, C. (1904). "General intelligence" objectively determined and measured. American Journal of Psychology, 15, 201-293.

Sternberg, R. J. (1983). Components of individual differences in human intelligence (Final Report NR 150-412). Arlington, VA: Office of Naval Research.

Sternberg, R. J. (1985). Human abilities: An information processing approach. New York: Freeman and Co.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. Journal of Applied Psychology, 52, 372-376.

White, M. J. (1985). On the status of cognitive psychology. American Psychologist, 40, 117-119.

Appendix A

Copy of the Instructions for

the Flight Planning Test

pages 85-95

inclusive

# F L I G H T   P L A N N I N G   T E S T

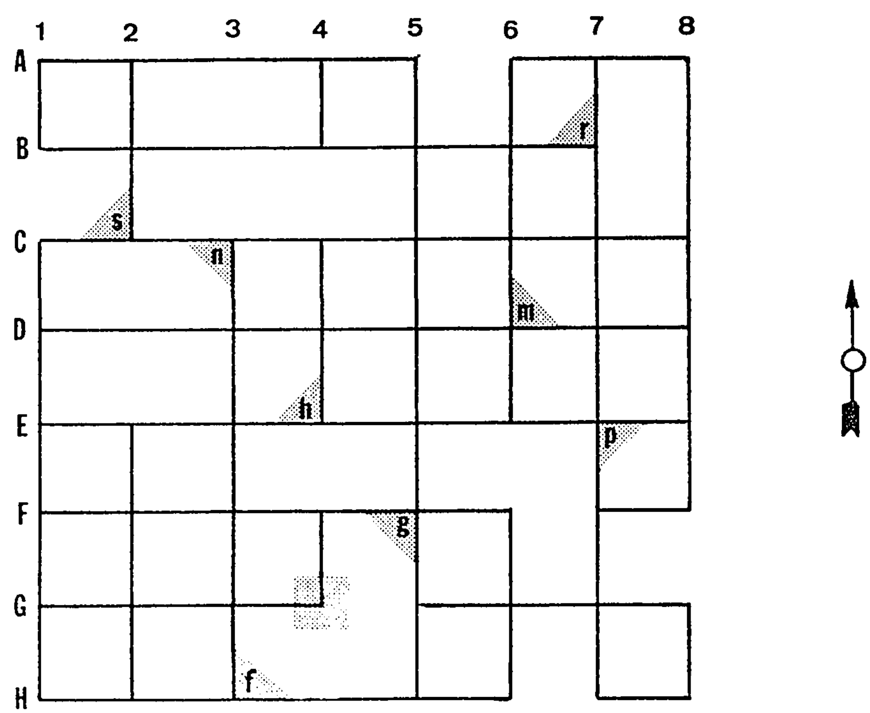
## GENERAL DIRECTIONS

This is a test of your ability to use several types of information to plan a simulated helicopter flight. On each part of the test, you will be shown a route map and then asked questions about 12 flights. Use the following directions to answer each question.

Look at the practice route map below. The numbers across the top and letters down the left side are map coordinates. Any map intersection can be identified by a letter-number label. For example, the upper left corner of the map is A1 and the lower right corner is H8. The airfield (▒) is located at G4.

Flights can only be planned along approved air routes indicated by solid lines. The lettered triangles at the intersection of some air routes are landmarks. For example, the landmark m is at D6. The map also has an arrow pointing in a North (N) direction. The other compass headings are East (E), South (S), and West (W).

## PRACTICE ROUTE MAP



On each test item, you will have to determine the "best" route between two intersections by applying the following rules:

1. The "best" route is always the shortest route.
2. If two or more routes are the same length, the "best" route has the fewest turns (a turn is any change in direction after the initial direction).
3. If two or more routes are the same length and have the same number of turns, the "best" route passes the most landmarks (the route must pass along the side of the triangle; the landmark point cannot only begin or end the route).

For example, the best route from D3 to F5 is via F3, which passes landmark g. The route from D3 to D5 to F5 is the same distance and number of turns, but does not pass a landmark.

## SAMPLE ITEMS

Use the practice map to answer the sample items below. Four answer choices will follow each item. Decide on the best answer, then mark the same letter on the answer sheet. You will have 2 minutes to answer the sample items.

S1. Which landmark(s) would you pass in flying from B6 to D8?

A. m B. mp C. p D. r

S2. How many turns would be needed to fly from D2 to G5?

A. None B. 1 C. 2 D. 3

S3. What compass headings would you fly from G2 to A8?

A. N,E,N,E B. S,E,N C. W,N,E D. N,E,N

STOP

DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO!

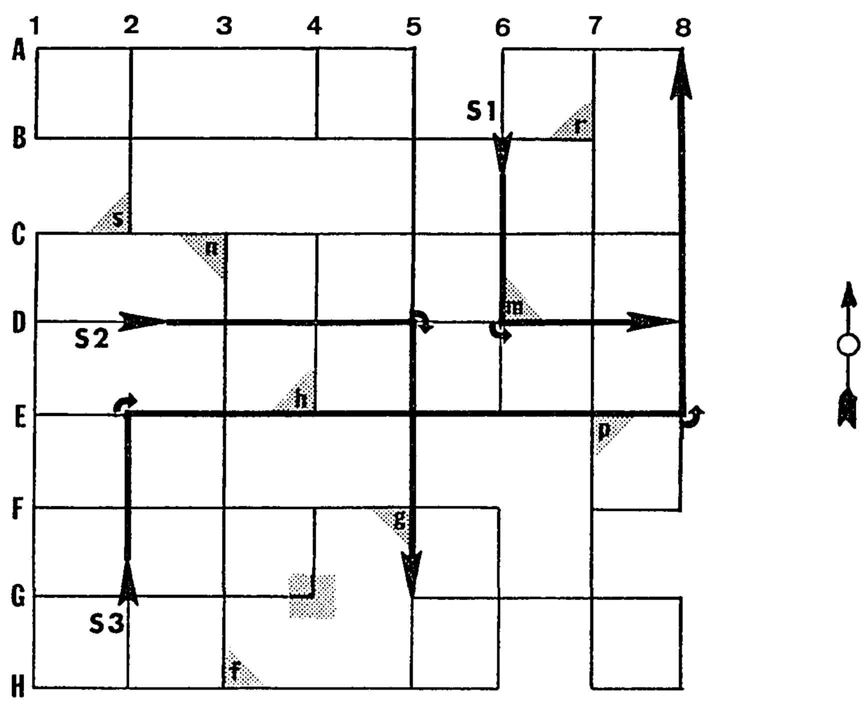
The answers to the sample questions are illustrated below.

S1. The answer is A. The best route (line S1) passes landmark m. You could also pass by landmark r, but that route would require two turns instead of one.

S2. The answer is B. The best route (line S2) requires one turn. Any other route would either be longer or require more turns.

S3. The answer is D. The best route (line S3) follows N, E, N headings. Any other route would either be longer or require more turns.

**PRACTICE ROUTE MAP**




GO ON TO THE NEXT PAGE

Before proceeding, you will have 30 seconds to review the basic rules of flight planning:

- Solid lines are approved air routes.
- Each route intersection is identified by a letter-number label.
- The arrow points toward the North.
- The lettered triangles indicate landmarks.
- To pass a landmark, the route must run along the side of the triangle.
- A turn is any change of direction after the initial direction.
- In order of importance, select the air route that:
  1) requires the shortest distance;
  2) requires the fewest turns;
  3) passes the most landmarks.

## ROUTE MAP 1S

You will have 9 minutes to answer the 12 items on Route Map 1S. If you cannot determine the answer to an item but can eliminate some of the choices, make your best guess. If you change an answer, erase your first answer completely. Do not mark on the test booklet.

If you finish Route Map 1S, check your answers. Do not go on to the next route map or turn back to the instructions.

STOP

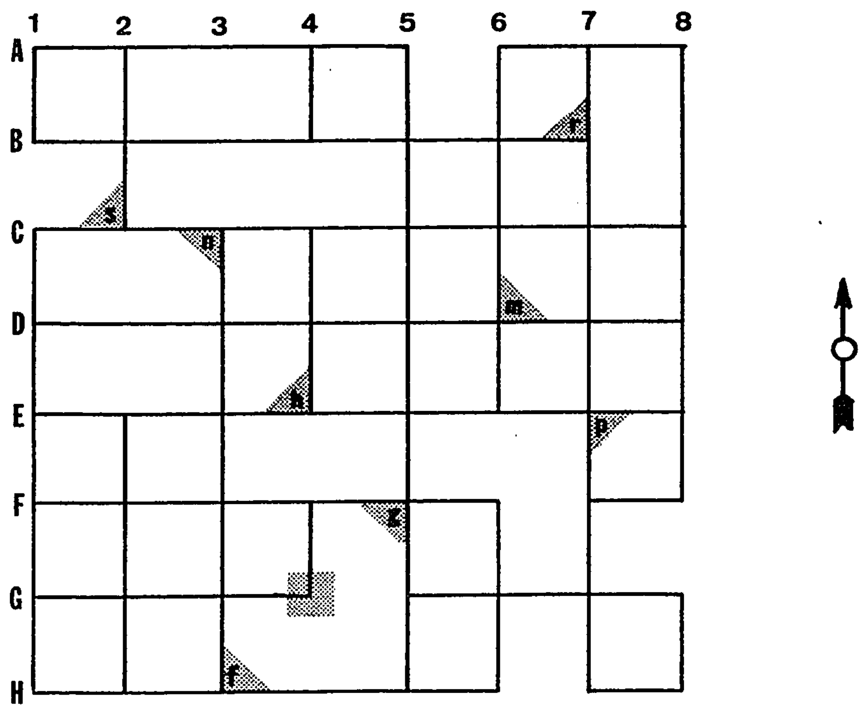
DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO!

## LEVEL 2 DIRECTIONS

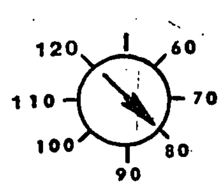
In addition to the general directions, the following information may be used to answer items on the remaining route maps:

1. The vertical or horizontal distance between adjacent map intersections is 1Ø miles. On the practice map below, the distance from A1 to B1 is 1Ø miles; from A1 to B2 2Ø miles.

**PRACTICE ROUTE MAP**




2. Airspeed may be presented as numbers in miles per hour (e.g., 8Ø mph), or shown on an airspeed indicator. In the example below, the arrow shows an airspeed of 8Ø mph. The mph values will not be labeled on the indicator during the test. Memorize the mph values on the indicator.




GO ON TO THE NEXT PAGE

To answer items about the required airspeed or flight time, determine the distance of the best route in miles, the flight time in minutes, and the airspeed in mph. Assume a constant airspeed over the entire route.

If the time is given, the required airspeed is equal to:

**(miles ÷ minutes) x 60**

If a route covers 90 miles and the flight time is 45 minutes:

Airspeed = (90 miles ÷ 45 minutes) x 60 = 2 x 60 = 120 mph

If the airspeed is given, the required time is equal to:

**(miles ÷ airspeed) x 60**

If a route covers 105 miles and the airspeed is 75 mph:

Time = (105 miles ÷ 75 mph) x 60 = 1.4 x 60 = 84 minutes

The required time or airspeed usually can be determined mentally, but scratch paper is provided, if required. Do not copy the formulas or mph values on the scratch paper.

## SAMPLE ITEMS

Use the practice map to answer the sample items below. Decide on the best answer, then mark the same letter on the answer sheet. You will have 3 minutes to answer the sample items.

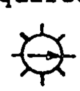
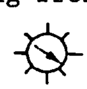
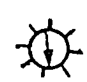
S4. How many minutes are required in flying from E8 to G4 at an airspeed of 120 mph?

A. 20 B. 30 C. 40 D. 50

S5. How many minutes are required in flying from H1 to D6 at an airspeed of  ?

A. 24 B. 36 C. 48 D. 54

S6. What airspeed is required in flying from B8 to A5 in 46 minutes?

A.  B.  C.  D. 

STOP

DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO!

The answers to the sample items are illustrated below.

S4. The answer is B (30 minutes). The best route (line S4) is 60 miles in length. The flight time is equal to:

(60 miles ÷ 120 mph) x 60 = .5 x 60 = 30 minutes

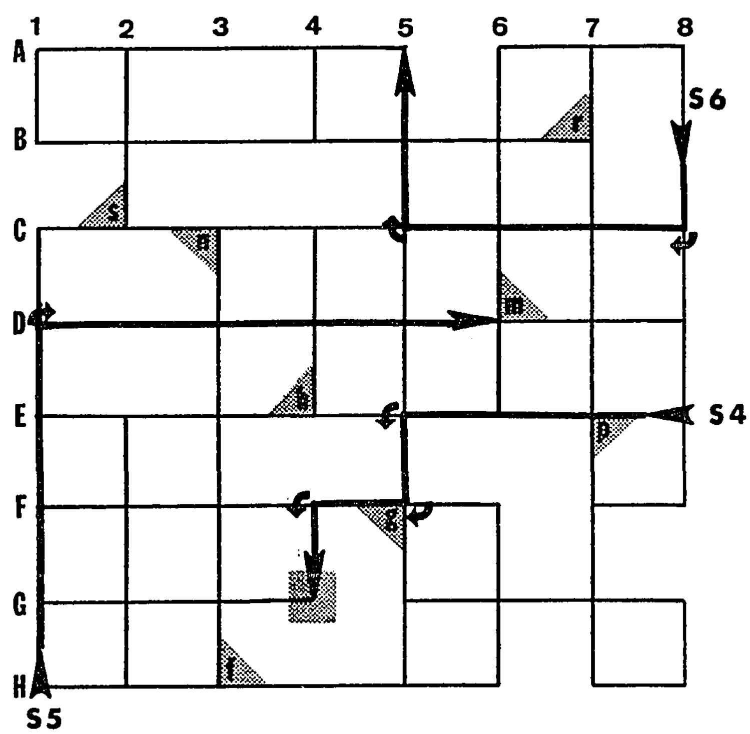
S5. The answer is D (54 minutes). The best route (line S5) is 90 miles in length. The flight time is equal to:

(90 miles ÷ 100 mph) x 60 = .9 x 60 = 54 minutes

S6. The answer is C (78.24 mph or  ). The best route (line S6) is 60 miles in length. The airspeed is equal to:

(60 miles ÷ 46 minutes) x 60 = 1.304 x 60 = 78.24 mph

**PRACTICE ROUTE MAP**




GO ON TO THE NEXT PAGE

On item S6, a precise speed of 78.24 mph is not necessary since the choices differ by 8 mph. Also, an airspeed of 80 mph could be estimated by rounding off to 45 minutes. Answer C is the closest value to this estimate.

Before proceeding, you will have 30 seconds to review the Level 2 (L2) rules:

- Solid lines indicate approved air routes.
- Each route intersection is identified by a letter-number label.
- The arrow points toward the North.
- The lettered triangles indicate landmarks.
- To pass a landmark, the route must pass a side of the triangle.
- A turn is any change of direction after the initial direction.
- Select the air route that: 1) requires the shortest distance, 2) requires the fewest turns, and 3) passes the most landmarks.
- The vertical or horizontal distance between adjacent route intersections is 1∅ miles.
- The airspeed indicator increases clockwise in intervals of 1∅ miles per hour (mph), from 6∅ mph (upper-right) to 12∅ mph (upper-left).
- The flight time formula is: time = (miles ÷ airspeed) x 6∅.
- The airspeed formula is: airspeed = (miles ÷ minutes) x 6∅.

You will have 9 minutes to answer the 12 items on the following route map. Work as rapidly as possible without sacrificing accuracy. If you finish the route map, check your answers.

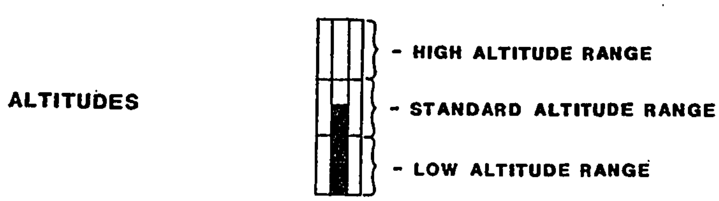
STOP

DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO!

## LEVEL 3 DIRECTIONS

In addition to the previous directions, the following information may be used to answer items on the remaining route maps:

1. Flight altitude may be presented verbally as high, standard, or low altitude, or shown as a range by the dark line on an altimeter. The altimeter below is at standard altitude:




Memorize the ranges (high, standard, or low) on the altimeter; the altimeter ranges will not be labeled during the test.

2. Flight restrictions between two adjacent points are indicated by circles placed on approved air routes.

   - A mostly dark circle (  ) prohibits standard and low altitude flights, but permits high altitude flights.
   - A partly dark circle (  ) prohibits low altitude flights, but permits standard and high altitude flights.

Memorize the restrictions indicated by each symbol. The circles will not be labeled on the map routes.

GO ON TO THE NEXT PAGE

Before proceeding, you will have 30 seconds to review the Level 3 (L3) rules:

- Solid lines indicate approved air routes.
- Each route intersection is identified by a letter-number label.
- The arrow points toward the North.
- The lettered triangles indicate landmarks.
- To pass a landmark, the route must pass a side of the triangle.
- A turn is any change of direction after the initial direction.
- Select the air route that: 1) requires the <u>shortest distance</u>, 2) requires the <u>fewest turns</u>, and 3) passes the <u>most landmarks</u>.
- The vertical or horizontal distance between adjacent route intersections is 10 miles.
- The airspeed indicator increases clockwise in intervals of 10 mph, from 60 mph (upper-right) to 120 mph (upper-left).
- The flight time formula is: time = (miles ÷ airspeed) x 60.
- The airspeed formula is: airspeed = (miles ÷ minutes) x 60.
- Flight altitude may be in the low, standard, or high range.
- Low altitude is below the first altimeter line; high altitude is above the second altimeter line.
- A partly dark ( ◒ ) circle placed on an approved air route prohibits low altitude flights between the adjacent points.
- A mostly dark ( ● ) circle placed on an approved air route prohibits low and standard altitude flights between the adjacent points.

You will have 9 minutes to answer the 12 items on the following route map. Work as rapidly as possible without sacrificing accuracy. If you finish the route map, check your answers.

STOP

**DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO.**

Appendix B

Copy of the Instructions for
the Chart Use Test

pages 97-99
inclusive

# CHART USE TEST

## DIRECTIONS

This is a test of your ability to follow directions in using charts. No previous experience with the test charts is needed to accomplish the required steps. Do not assume any additional information or be concerned about the meaning of the variables or abbreviations.

For each chart, you will first be shown two examples of its use. Some chart items will be exactly like the examples, some items will be similar but reversed, and other items may require additional steps or use different variables. However, sufficient information is presented in the examples to complete all the items on each chart.

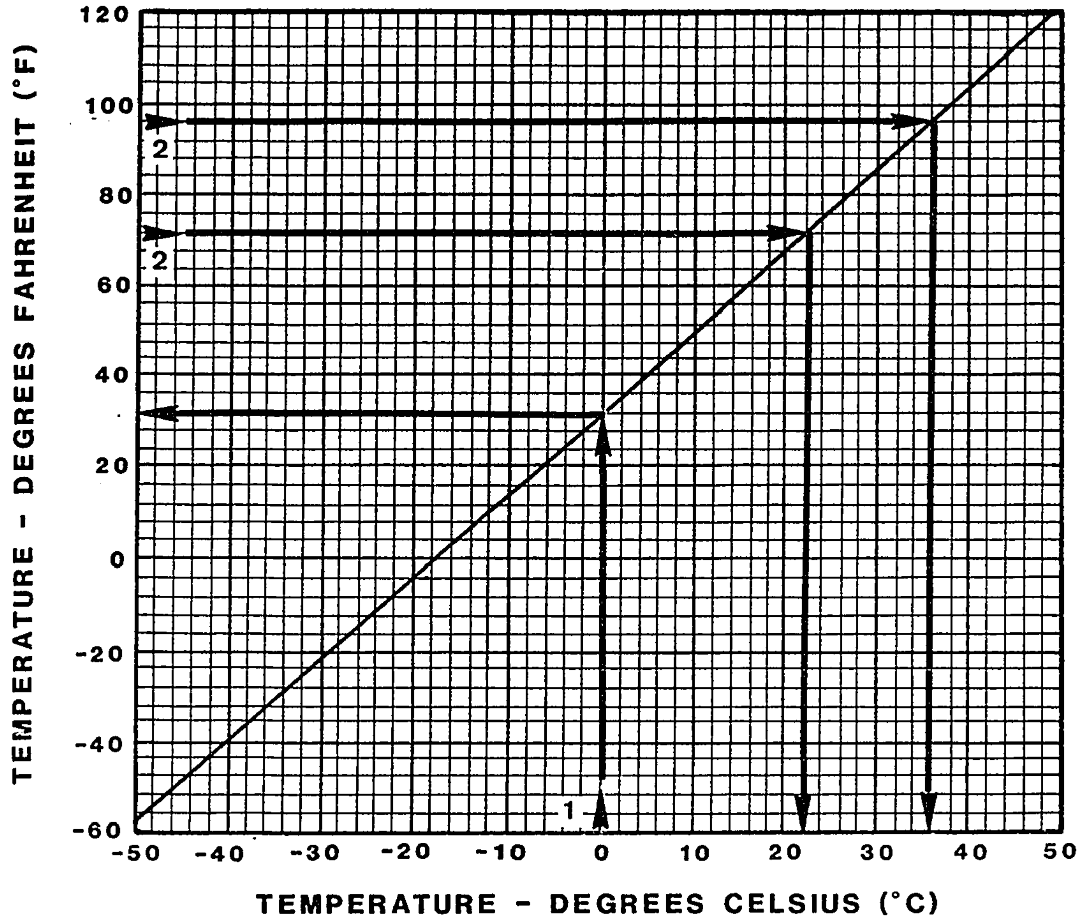
On each chart item, use the "known" information and the chart to determine the "wanted" information. Four answer choices will follow each item. Select the best choice and mark the corresponding letter on the answer sheet.

The test is divided into five, separately timed, charts with six items each. You will have two minutes to review the two examples on using each chart. Then you will have four minutes to work the six items. Work as rapidly and accurately as possible.

If you finish all the items on a chart, check your work. Do not go on or go back to other charts unless you are instructed to do so.

GO ON TO THE NEXT PAGE

## TEMPERATURE CONVERSION CHART




EXAMPLE 1. Known: Celsius temperature = 0°C
Wanted: Fahrenheit temperature (°F)

A. 28 B. 32 C. 36 D. 40

METHOD: Enter Celsius scale at 0°C and move up to the diagonal line
Move left to Fahrenheit scale and read 32°F
Darken choice B on the answer sheet for EX1.

EXAMPLE 2. Known: 96°F, 72°F
Wanted: difference in °C

A. 6 B. 10 C. 14 D. 24

METHOD: Use chart to find 96°F = 36°C
Use chart to find 72°F = 22°C
Subtract to find the difference: 36 - 22 = 14°C
Darken choice C on the answer sheet for EX2.

STOP

DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO

Appendix C

Copy of the Instructions for
the Map Planning Test

pages 99-101
inclusive

## MAP PLANNING TEST

### DIRECTIONS

This is a test of your ability to plan a route between two points. You will be shown sections of city maps where some streets are blocked. Your task will be to find the shortest passable route between each set of points as quickly as possible.

Look at the sample map on the next page. The solid lines are city streets. The circles show places where the streets are blocked. You will have to plan routes between the points indicated by the letters on the edge of the map.

The numbered squares are buildings. The shortest route between two points will pass along the side of one, and only one, of the numbered buildings. A route must run along the side of a building to pass it. A building is not passed if the route only touches the corner.

On each item, you must determine the shortest route between the two points. Mark the corresponding letter on your answer sheet to indicate the building number that is passed on your chosen route. Use the sample map to answer practice item P1.

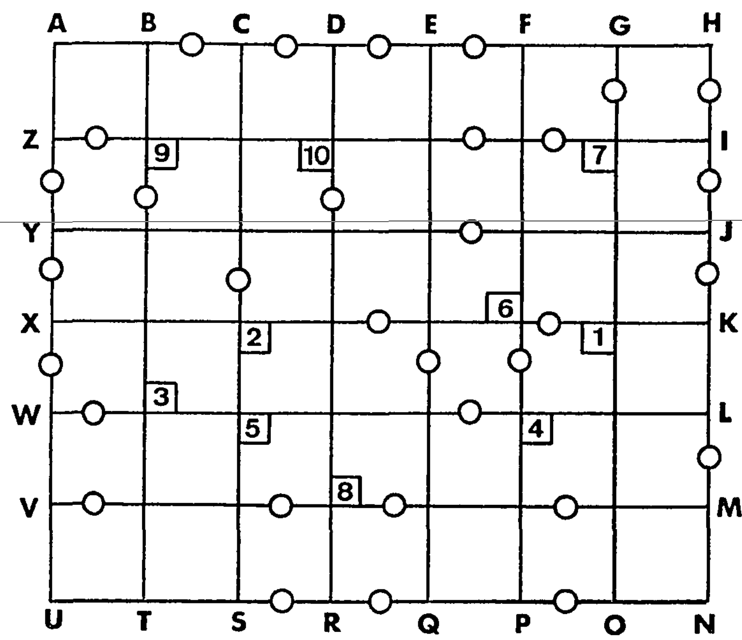
P1. The shortest route from K to I passes building:

A. 1 B. 4 C. 6 D. 7

The shortest route passes building 7, so mark letter D on the answer sheet for P1. Building 1 (letter A) is not correct because the route only touches the corner of the building, not a side. The shortest route will pass only one building.

Use the sample map to work practice items P2 through P5. Darken the letter on the answer sheet for the building that is passed on each route. You will have one minute to work practice problems P2-P5.

GO ON TO THE NEXT PAGE




| | THE SHORTEST ROUTE FROM | PASSES BUILDING | | | |
|---|---|---|---|---|---|
| P2. | P to S | A. 3 | B. 4 | C. 5 | D. 8 |
| P3. | A to Q | A. 2 | B. 5 | C. 8 | D. 9 |
| P4. | V to C | A. 2 | B. 3 | C. 5 | D. 9 |
| P5. | Y to J | A. 6 | B. 7 | C. 9 | D. 10 |

The practice routes pass the following buildings: P2 passes 5(C); P3 passes 9(D); P4 passes 3(B); P5 passes 6(A).

The test is divided into three parts of 20 items each (2 maps on each part). You will have three minutes to work on each part of the test. Your score on this test will be the number of correct answers. Work as rapidly as possible without sacrificing accuracy.

When you complete a part of the test, check your answers. Do not go back to a previous part, or turn to the next part until told to do so. Please do not mark on the city maps.

S T O P

DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO!

Appendix D

Copy of the Instructions for
the Sound Reasoning Test

pages 102-104
inclusive

## SOUND REASONING TEST

### DIRECTIONS

This is a test of your ability to draw a correct conclusion using only the information that is readily available. On each item, two statements will be given followed by a choice of four conclusions. Although the terms used in the statements are nonsense words, assume that the statements are true. Select the correct conclusion solely on the basis of sound reasoning, then mark the corresponding letter on the answer sheet. Look at the following practice item:

P1. All COR are LIM
All FAX are COR
Therefore,

A. All FAX are LIM

B. All LIM are FAX

C. No LIM are FAX

D. All FAX are not LIM

The correct answer is A. Even though the terms are nonsensical, if all FAX are COR and all COR are LIM, it is sound reasoning that all FAX are LIM also.

The other choices are not sound reasoning. There could be other LIM that are not COR or other COR that are not FAX, so choice B is not correct. The statements also imply that at least some LIM must be FAX, so choice C is not correct. Finally, since all FAX are LIM is sound reasoning, choice D cannot be sound reasoning. Mark letter A for item P1 on the answer sheet.

GO ON TO THE NEXT PAGE

You will have one minute to work the following two practice items. Decide which conclusion reflects sound reasoning, then mark the corresponding letter for items P2 and P3 on the answer sheet.

P2. No TEL are PAQ
All MIT are PAQ
Therefore,

A. Some TEL are MIT

B. All TEL are MIT

C. All MIT are TEL

D. No MIT are TEL

P3. All RUF are SEP
Some LOH are RUF
Therefore,

A. No LOH are SEP

B. Some LOH are SEP

C. All SEP are LOH

D. Some SEP are not LOH

The correct answer for P2 is D. If no TEL are PAQ but all MIT are PAQ, MIT and TEL must be completely exclusive.

The correct answer for P3 is B. Since some LOH are RUF and all RUF are SEP, at least some LOH must be SEP. Answers C and D on P3 may be true also, but they do not necessarily follow from the statements.

The test is divided into two parts of 16 items each. You will have six minutes to work on each part of the test. Work as rapidly as possible without sacrificing accuracy.

If you complete the first part of the test, check your answers. Do not turn to the second part until told to do so.

STOP

DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO!

Appendix E

Copy of the Instructions for

the Finding Rules Test

pages 105-107

inclusive

# FINDING RULES TEST

## DIRECTIONS

This is a test of your ability to find and use a rule or set of rules to solve a problem. Each item in this test consists of five lines containing 20 elements per line. An element is a dot, a circle, or a blank that separates the other elements into groups. A single dot is considered a group of one.

Each of the first four lines has only one circle. It has been placed on the line according to a rule or set of rules. You must determine the rule(s) used to place the circle in the first four lines, then apply the rule(s) to the fifth line. Decide which lettered dot should be the circle if the same rule(s) are used, then mark the same letter on the answer sheet.

Look at the following sample item:

```
... ..... ...o .....
..... .... ......o .
... ... ...o .. ....
. .. ..o .... ... ..

... . ... .... .. ..
  A     B   C   D
```

Two rules are required to place the circle on the fifth line. The circle is always the last element in a group, but this rule applies to the elements lettered A, B, and D. The circle is also always in the third group from the left end. The element lettered B is in the third group from the left end. Therefore, letter B would be marked on the answer sheet.

GO ON TO THE NEXT PAGE

Examples of the types of rules are: the location of the circle in a group (e.g., first, third, odd, or even element), the group size (e.g., one, two, odd, or even), and the size or location of groups that precede or follow the circle. There will always be at least one rule but not more than three rules used to place the circle on a line.

You will have one minute to work the two practice items that follow. Mark your answer sheet for items Pl and P2. If you complete the practice items, check your work.

```
P1.  ..... .o. ... ......
     .o. .... ..... ... .
     .... .o. .... ... ..
     .. ..... .... .o. ..

     ..... ... ... ... ..
       A    B    C    D
```

```
P2.  .... ... ..... . .o.
     . .o.. .. ..... ....
     .. .. . o... .... ..
     ..... . ...o ... ...

     .. . .. ... .... ...
          A    B    C    D
```

In Item Pl, the correct answer is C. The circle must be the middle element in a group of three. In Item P2, the correct answer is A. The circle is any element in a group that immediately follows a single dot.

The test is divided into two parts of 16 items each. You will have eight minutes to work on each part. Work as rapidly and accurately as possible. When you have finished Part I, check your work. Do not go on to Part II until you are instructed to do so.

STOP

DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO!

Appendix F

Copy of the Instructions for

the Rapid Match Test

pages 108-110

inclusive

# RAPID MATCH TEST

## DIRECTIONS

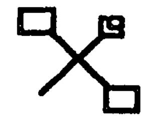
This is a test of your ability to determine as rapidly as possible if figures are the same as or different from a target figure. Look at the following example of a television set. The figure on the far left is the target that is to be compared to the four alternative figures. Alternative A is identical to the target figure. Alternatives B, C, and D are different from the target: The picture tube is a different shape, the dials are in a different position, and the antenna is broken, respectively.

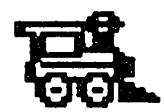
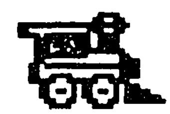
EX.  A.  B.  C.  D. 

The test is divided into two sections. On the IDENTICAL section, determine which alternative is the same as the target figure, then mark the corresponding letter on the answer sheet. On the DIFFERENT section, determine which alternative is different from the target figure, then mark the corresponding letter on the answer sheet.

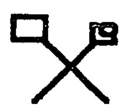
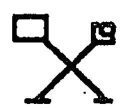
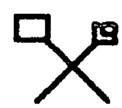
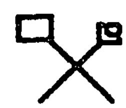
You will have 30 seconds to work the practice problems on the following page. Mark the corresponding letter on the answer sheet for items P1 through P4. If you complete all the items, check your answers. Do not read ahead.

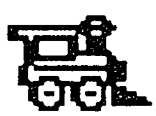
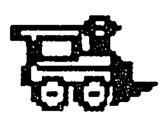
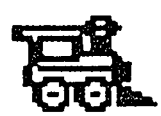
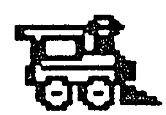
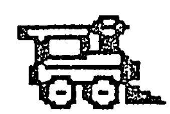
## RAPID MATCH TEST--IDENTICAL

P1.  A.  B.  C.  D. 

P2.  A.  B.  C.  D. 

## RAPID MATCH TEST--DIFFERENT

P3.  A.  B.  C.  D. 

P4.  A.  B.  C.  D. 

The correct answers for P1 and P2 are A and C, respectively. The other alternatives are slightly different from the respective target figures. These are examples of the IDENTICAL section.

The correct answers for P3 and P4 are B and A, respectively. The other alternatives are identical to the respective target figures. These are examples of the DIFFERENT section.

You will have four minutes to complete each section of the test. Work as rapidly as possible without sacrificing accuracy. If you complete the first section, check your answers. Do not turn to the next section until told to do so.

STOP

DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO.

Appendix G

Copy of the Instructions for
the Figure Orientation Test

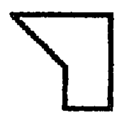
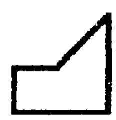
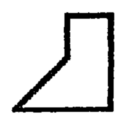
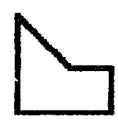
pages 111-113
inclusive

# FIGURE ORIENTATION TEST

## DIRECTIONS

This is a test of your ability to determine if figures are the same as a target or are different from a target despite changes in orientation. A figure is different from the target if it is inverted (i.e., flipped over). Look at the five figures below.

The figure on the far left is the target that is to be compared to the four alternative figures. Alternatives A and B are the same figure except that they are rotated to a different position. Alternatives C and D are inverted and then rotated; they are not the same as the target figure.

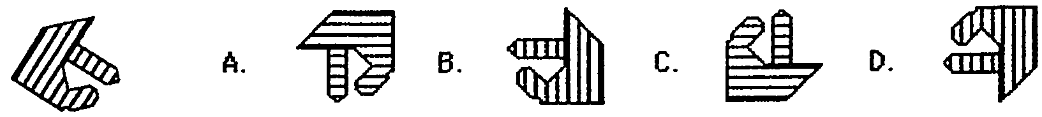
EX.  A.  B.  C.  D. 

The test is divided into two sections. On the ROTATED section, determine which alternative is the same as the target figure except for the change in orientation, then mark the corresponding letter on the answer sheet. On the INVERTED section, determine which alternative is inverted as well as rotated, then mark the corresponding letter on the answer sheet.

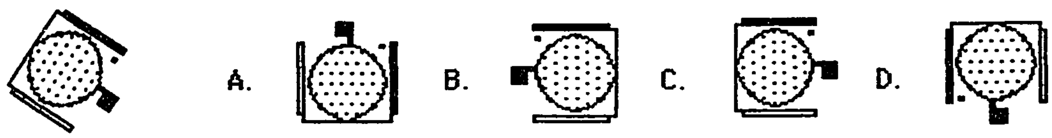
You will have 30 seconds to work the practice problems on the following page. Mark the corresponding letter on the answer sheet for items P1 through P4.

ORIENTATION TEST--ROTATED

P1.

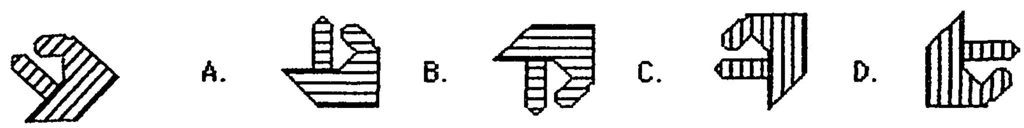



P2.

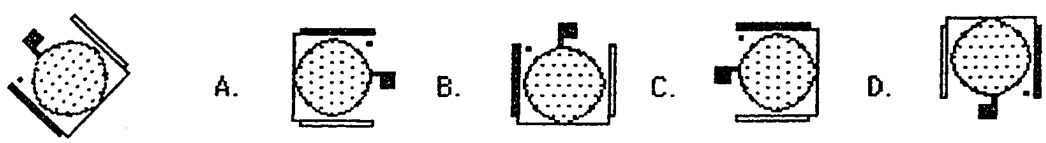



ORIENTATION TEST--INVERTED

P3.




P4.




The correct answers for P1 and P2 are D and C, respectively. These are examples of the ROTATED section. The correct choices are rotated forms of the target figures. The other choices are inverted as well as rotated.

The correct answers for P3 and P4 are B and C, respectively. These are examples of the INVERTED section. The correct choices are inverted and rotated forms of the target figures. The other choices are only rotated.

You will have five minutes to work on each section of the test. Work as rapidly as possible without sacrificing accuracy.If you complete the first section, check your answers. Do not turn to the next section until told to do so.

STOP

DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO.

Appendix H

Copy of the Instructions for
the Finding Figures Test

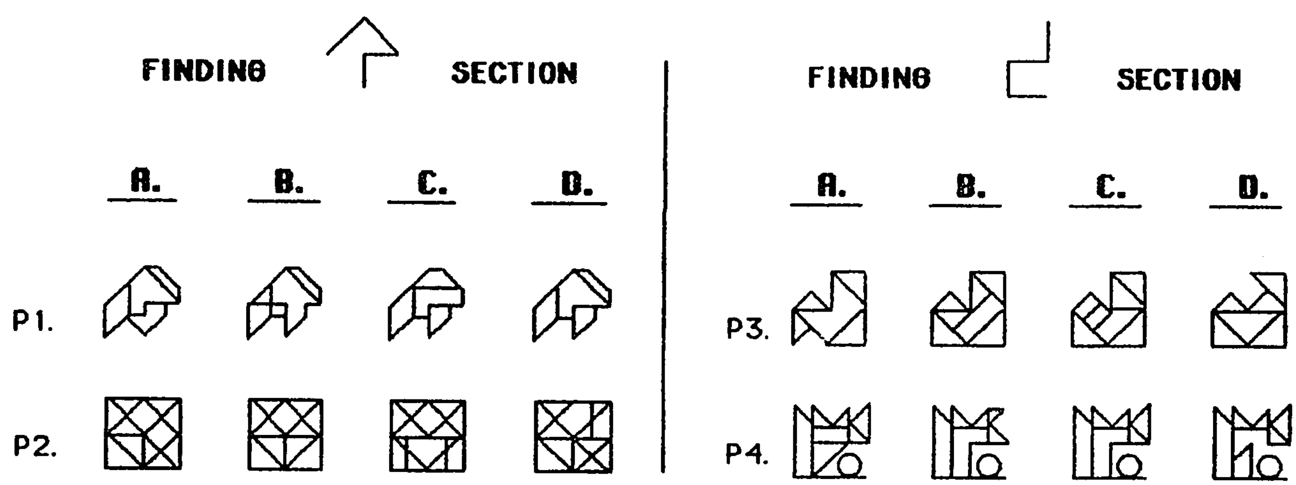
pages 114-116
inclusive

# FINDING FIGURES TEST

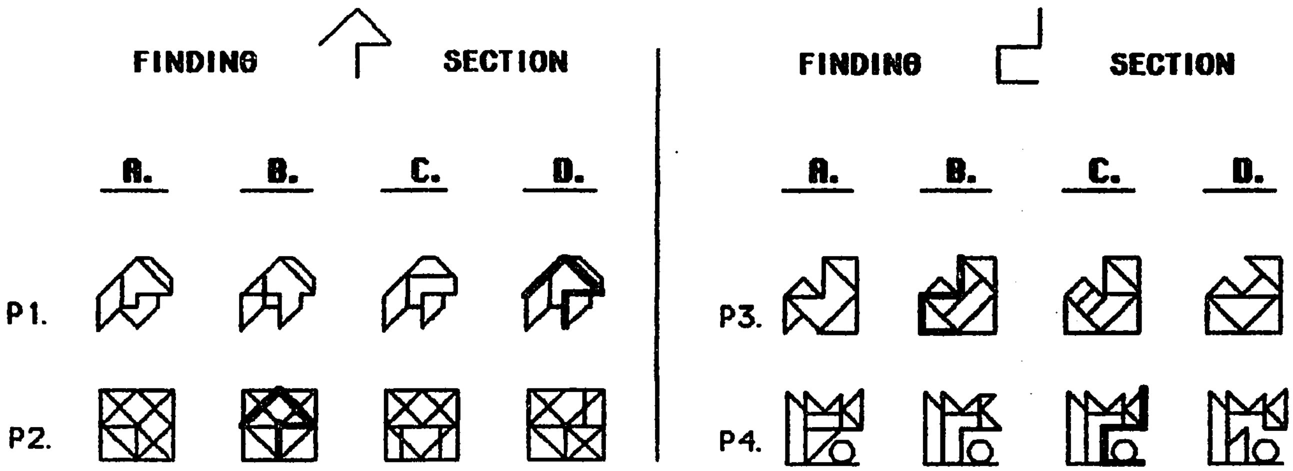
## DIRECTIONS

This is a test of your ability to recognize as quickly as possible a figure that is hidden in a larger pattern. The target figure will be shown at the top of each section. On each item, there will be four patterns under the columns A, B, C, and D. The target figure will be hidden in one of the four patterns. The figure will always be in the same orientation. That is, the figure will not be rotated or inverted.

You will have 30 seconds to answer the practice items below. Decide which pattern contains the entire target figure, then mark the corresponding letter on the answer sheets.




The hidden figures are illustrated on the following page. The correct answers for P1 and P2 are D and B, respectively. The patterns under the other columns do not contain the complete target figure. The correct answers for P3 and P4 are B and C, respectively. Only B and C contain the complete target figure.




There are two, separately timed, sections on this test. On one section, you will be searching for the target figure in P1 and P2. On the other section, you will be searching for the target figure in P3 and P4. Be sure to mark your answers on the correct sheet.

You will have five minutes to work on each section of the test. Work as rapidly as possible without sacrificing accuracy. If you complete the first section, check your answers. Do not turn to the next section until told to do so.

STOP

DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO.

Appendix I

Copy of the Instructions for
the Obscured Figures Test

pages 117-119
inclusive

## OBSCURED FIGURES TEST

### DIRECTIONS

This is a test of your ability to identify a figure that is partially obscured. On each item, an obscured figure and four alternative figures will be presented. Decide which alternative best represents the obscured figure, then mark the corresponding letter on the answer sheet. If you cannot identify the obscured figure but can eliminate some of the alternatives, make your best guess.

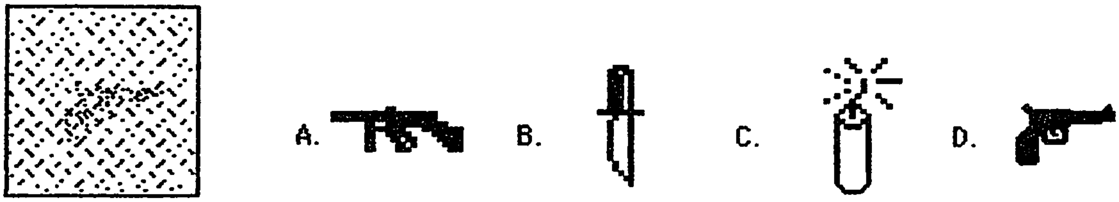
There are two sections on this test. On one section, the obscured figure will be presented on a white background. On the other section, the obscured figure will be presented on a black background. The figures will always be in the same orientation. That is, the figures will not be rotated or inverted. Look at the examples on the following page.

The correct answers for examples 1 and 2 are D (pistol) and C (knife), respectively. These are examples of the white background figures. The correct answers for examples 3 and 4 are A (machine gun) and B (dynamite), respectively. These are examples of the black background figures.

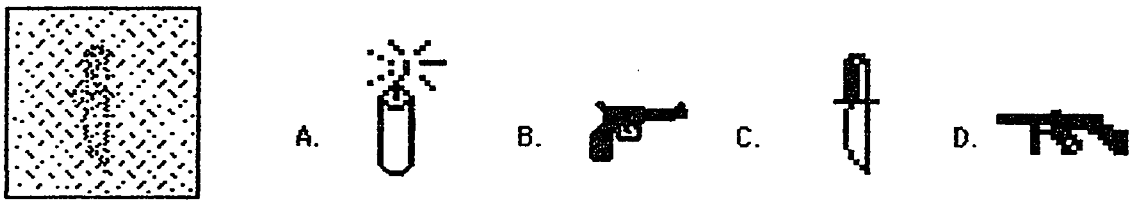
You will have two minutes to complete each section of the test. Work as rapidly as possible without sacrificing accuracy. If you complete the first section, check your answers. Do not turn to the next section until told to do so.

GO ON TO THE NEXT PAGE

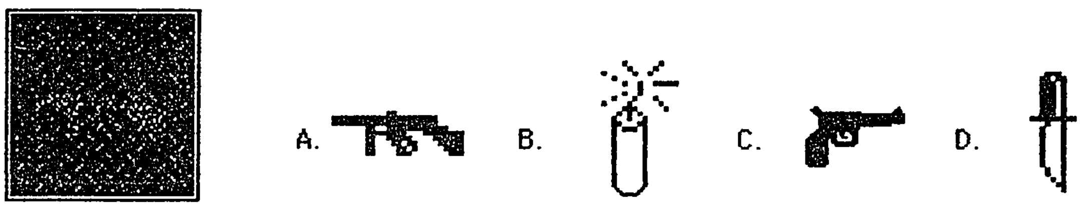
EX 1.



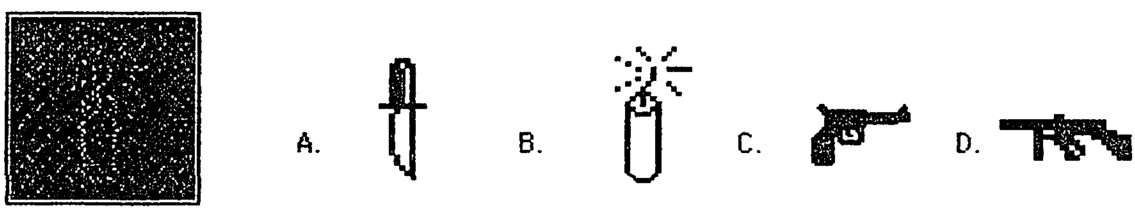

EX 2.




EX 3.




EX 4.




STOP

DO NOT GO ON TO THE NEXT PAGE UNTIL TOLD TO DO SO.

Appendix J

Design Specifications for the

Flight Planning Test

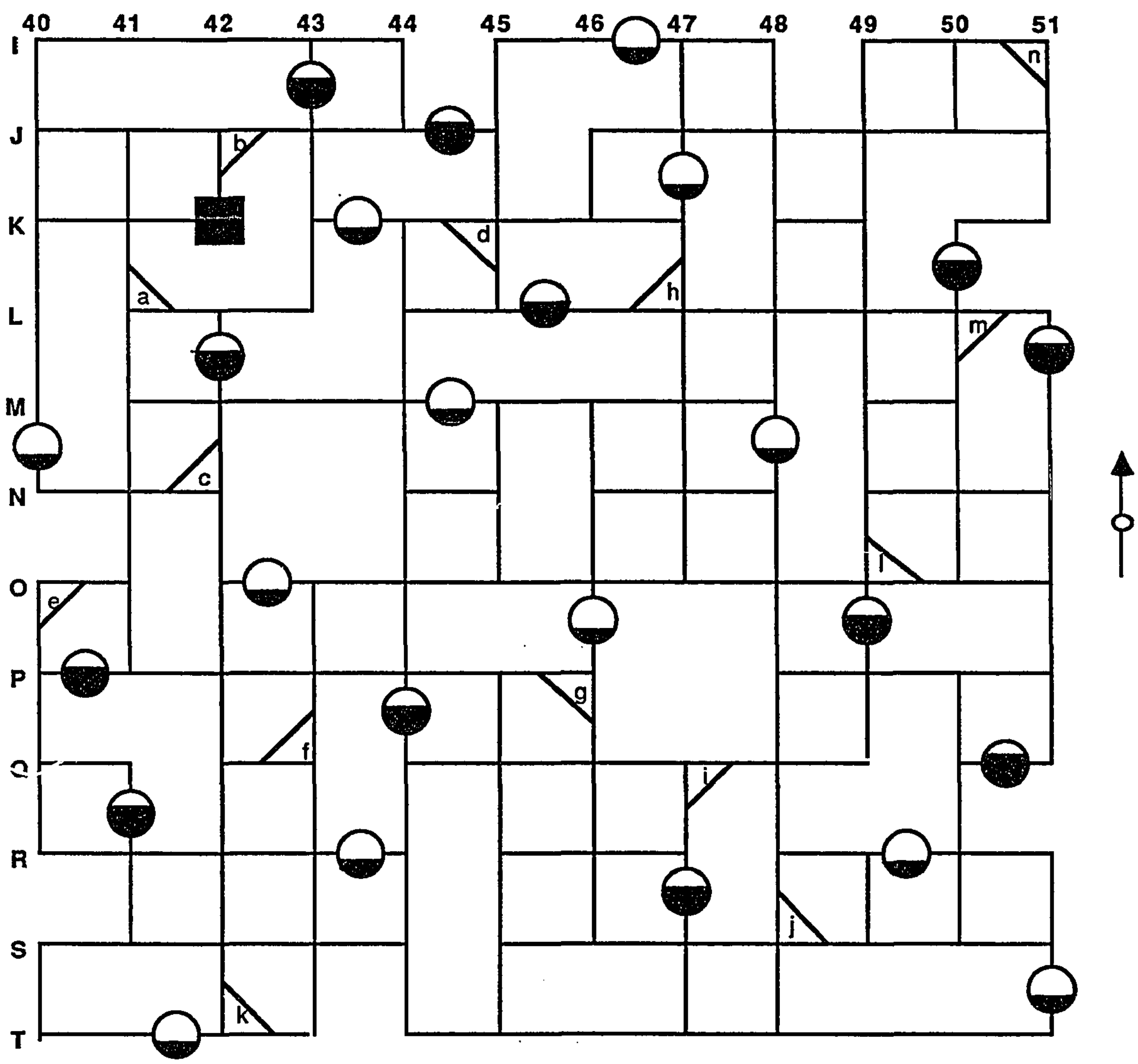
pages 120-129

inclusive

FLIGHT PLANNING TEST

Route Map Specifications

Seven route maps representing three levels of complexity (one basic, three intermediate, and three advanced maps) will be developed for the Flight Planning Test. General specifications that apply to all maps are presented first, followed by additional specifications for the successive complexity levels. The sample map below represents the advanced level but subsumes all lower level specifications.



General Specifications

Dimensions. The perimeter of each map will be square in shape. The exact dimensions will be specified for each level of complexity.

Compass Indicator. Approximately 2 cm outside the right perimeter (centered vertically), place an unlabeled arrow pointing toward the top of the page to indicate a northerly direction.

Coordinate Points. Beginning with the upper left corner and proceeding clockwise, locate a coordinate point every 1.7 cm on the perimeter. In the map interior, the intersection of every vertical and horizontal perimeter coordinate also constitutes a coordinate point.

Coordinate Labels. Beginning in the upper left corner, label the horizontal (column) coordinates across the top in numerically ascending order. Label the vertical (row) coordinates down the left side in alphabetically ascending order. The starting labels will be specified for each map. Each perimeter and interior coordinate point will then be described by a letter-number combination.

Air Routes. The horizontal and vertical link between adjacent points will be potential air routes. There will be no diagonal air routes. Approved flight paths will be indicated by a solid line drawn between adjacent points. If all routes were approved, each map would contain (row - 1) x (column - 1) square grids, each 1.27 $cm^2$ in area. Not all air routes should be approved. Lines may be removed to form:

- rectangles (limited to double or triple grid lengths or widths),
- L-shaped triple grids,
- square quadruple grids,

Guidelines for the grid proportions (single grids used in a shape divided by the total number of single grids in the map) are:

- single grids 40-50 percent,

- double rectangles 15-25 percent,
- triple rectangles 10-20 percent,
- triple L-shapes 10-20 percent, and
- quadruple squares 0-10 percent.

In addition, at least one, but not more than three, potential air routes should not be approved on each side of the perimeter. The placement of unapproved air routes is discussed further under item generation.

Airfield. A darkened, .635 x .635 cm square will indicate an airfield on each map to serve as a reference point. The airfield will be centered on an interior coordinate point and be connected to two approved air routes. The airfield location should differ on each map.

Landmarks. Map landmarks will be represented by right triangles. The number of landmarks will be specified for each map. Not more than two landmarks should be placed on any row or column. The vertex of the right triangle will be coincident with a right angle of a grid square. The length of each side will be .635 cm. Both sides of a landmark must abut an approved air route. Placement of landmarks is discussed further under item generation. Beginning at the airfield and extending outward in a counterclockwise rotation, label the landmarks with lower case letters in ascending order beginning with "a".

Difficulty Level 1 (DL1: Map 1) Specifications

Dimensions. The DL1 map will be 8.89 cm square. Each side of the map will contain eight coordinate points.

Coordinate Labels. Label the map 1 coordinates beginning with 1 and A.

Landmarks. The DL1 map will contain 10 landmarks.

Difficulty Level 2 (DL2: Maps 2, 3, 4) Specifications

Dimensions. The DL2 maps will be 11.43 cm square. Each side of the maps will contain 10 coordinate points.

Coordinate Labels. For maps 2, 3, and 4, respectively, label the coordinates beginning with 10, 20, and 30, and C, E, and G.

Landmarks. The DL2 maps will each contain 12 landmarks.

Difficulty Level 3 (DL3: Maps 5, 6, 7) Specifications

Dimensions. The DL3 maps will be 13.97 cm square. Each side of the maps will contain 12 coordinate points.

Coordinate Labels. For maps 5, 6, and 7, respectively, label the coordinates beginning with 40, 60, and 80, and I, L, and O.

Landmarks. The DL3 maps will each contain 14 landmarks.

Restrictions. DL3 maps will also contain 24 flight restrictions indicated by partially darkened circles, .40 cm in diameter, that are placed in the center of approved air routes. Twelve low-altitude restrictions will have the lower one-third of the circle darkened. Twelve standard-altitude restrictions will have the lower two-thirds of the circle darkened. Three low- and three standard-altitude restrictions will be located in each quadrant of the DL3 maps.

Item Generation

Twelve items will be generated for each map. All item stems will have the same general form:

- an interrogative (e.g., which, how many),
- a response term (e.g., landmarks, compass heading), and
- the flight coordinates (e.g., to fly from A1 to H8).

In addition, qualifying information needed to answer DL2 and DL3 questions, or intended to distract the examinee will be included on some item stems (e.g., at an airspeed of ____; at low altitude).

Selection of Routes (See Table 1)

The core question embedded in all items is the flight route that is preferred. That is, the direction, distance, heading, etc., depend on the route that is flown. Route preference is based on, in order of importance:

- the shortest route,
- the fewest turns, and
- the most landmarks passed on the route.

On each map, at least two of the items should be based on each preference criterion. Additional route parameters are discussed below.

Length. A route between adjacent coordinate points is a link, the underlying metric for length (on DL2 and DL3 maps, a link is equated to 10 miles). The frequency range for route lengths in each difficulty level will be:

| Difficulty | Length in Links (Minimum - Maximum) | | | |
|---|---|---|---|---|
| Level | 3-5 | 6-8 | 9-11 | 12-14 |
| DL1 | 1-3 | 4-6 | 3-5 | 0-1 |
| DL2 | 0-2 | 4-6 | 3-5 | 1-2 |
| DL3 | 0-1 | 3-5 | 4-6 | 2-3 |

Turns. The number of turns required on each flight will vary from one to six. At least 67 percent of the routes should require two, three, or four turns, in approximately equal proportions. Not more than two routes should require one turn, and not more than two routes should require five or six turns combined.

Landmarks. All flight routes which involve a landmark response, or depend on the number of landmarks as the preference criterion, must pass at least one landmark. Routes may pass a maximum of four landmarks.

Table 1

Route Selection Worksheet: Map# ____

| No. | Begin Pt. | End Pt. | Length | Turns | I-P | Dir | Criterion |
|---|---|---|---|---|---|---|---|
| 1 | | | | | P-P | D-R | |
| 2 | | | | | P-P | D-L | |
| 3 | | | | | P-P | U-R | |
| 4 | | | | | P-P | U-L | |
| 5 | | | | | P-I | D-R | |
| 6 | | | | | P-I | D-L | |
| 7 | | | | | P-I | U-R | |
| 8 | | | | | P-I | U-L | |
| 9 | | | | | I-P | D-R | |
| 10 | | | | | I-P | D-L | |
| 11 | | | | | I-P | U-R | |
| 12 | | | | | I-P | U-L | |

Horizontal Coordinates

___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

Begin

End

Vertical Coordinates

___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

Begin

End

Lengths

3 4 5 6 7 8 9 10 11 12 13 14

Turns

1 2 3 4 5 6

Route Terminals. The beginning and ending coordinate points on each flight route can be categorized as:

- perimeter to perimeter (P-P),
- perimeter to interior (P-I),
- interior to perimeter (I-P), and
- interior to interior (I-I).

All flight routes should have at least one terminal on the map perimeter; there should be no I-I routes. For each map, P-P, P-I, and I-P routes should each occur four times.

In addition, select starting and ending points so that vertical and horizontal coordinates are used in approximately equal proportions. The same coordinate point should not be used more than once as a beginning point or as an ending point on any map (i.e., not used more than twice on any map). In ordering items, the same horizontal or vertical coordinate should not be used to begin or end two successive flight routes. Similarly, the same coordinates should not end one route and begin the following route.

Flight Direction. Each route can be described as up (U) or down (D) and left (L) or right (R) from the beginning point to the ending point. The direction of flight should be equally distributed among the U-L, U-R, D-L, and D-R alternatives. Furthermore, there should be one route in each direction in each of the P-P, P-I, and I-P terminal categories on each map.

Selection of Stems

After selecting the 12 flight routes, an interrogative and response term must be selected. For the DL1 map, four routes should be randomly matched to each of the following stems:

(1) Which landmark(s) are passed in flying from ____ to ____?

(2) How many turns are required in flying from ____ to ____?

(3) What compass headings are required in flying from ____ to ____?

For the DL2 and DL3 maps, two routes should be randomly matched to each DL1 stem and to each of the following stems:

(4) How many miles are flown from ____ to ____?

(5) What airspeed is required in flying from ____ to ____ in ____ minutes?

(6) How many minutes are required in flying from ____ to ____ at an airspeed of ____?

Qualifying Information. Relevant information on time or airspeed is included for stems 5 and 6 on each DL2 and DL3 map. One of each stem 1-4 should include irrelevant information on time or airspeed. On DL3 maps, each stem should include qualifying information on flight altitude. At least two of the DL3 items should include altitude qualifications that are irrelevant or excessive.

Item Sequences. The order in which the items are presented will be randomly determined with the constraints of route terminal sequence (see page 6) and of having one of each stem presented before a stem is repeated.

Placement of Landmarks

When routes and stems have been selected, place landmarks on the map to identify routes for stem 1 items and to serve as a preference criterion where needed (see Selection of Routes). The remaining landmarks (see Map Specifications) should be distributed over the map to provide plausible response distractors and to balance the number of landmarks in each quadrant. The landmarks should be labeled according to the route map specifications.

Selection of Response Alternatives

All item responses will be in a five-alternative[1], multiple-choice format. One and only one alternative must be correct, but all alternatives should be feasible. If there is a logical order to the responses, alternatives A through E should be in ascending order. The correct answers should be approximately evenly distributed among the response alternatives A through E over all the maps. The use of "All (or none) of the above" and similar alternatives should be avoided, if possible.

[1] The response format was changed from five to four alternatives during the preliminary phase of testing. Based on the preliminary test data, the alternative that had the lowest frequency of selection was deleted. If two or more alternatives had an equally low selection rate, the alternative that was most different from the correct response was deleted.

# VITA

| | |
|---|---|
| Name: | Don Michael McAnulty |
| Born: | October 13, 1948<br>Raeford, NC |
| Personal: | Married to Beverly Styers McAnulty<br>Children: Rebecca, Timothy, and Jesse |
| Education: | Old Dominion University<br>Norfolk, VA |
| 1986 | Doctor of Philosophy<br>Industrial/Organizational Psychology<br>Director: Dr. Glynn D. Coates<br>Topic: "The development of a paper-and-pencil measure of complex cognitive-perceptual aptitude" |
| 1980 | Master of Science<br>Experimental Psychology<br>Director: Dr. Raymond H. Kirby<br>Topic: "Figural identity effects on the perception of apparent movement produced by a bistable multielement motion display" |
| 1978 | Bachelor of Science<br>Psychology |

Experience:

| | |
|---|---|
| 1983-Present | Senior Research Psychologist<br>Anacapa Sciences, Inc.<br>Fort Rucker, AL |

| | |
|---|---|
| 1982-1983 | Training and Development Coordinator<br>Human Resources Department<br>Farm Fresh Supermarkets<br>Norfolk, VA |
| 1980-1983 | Graduate Teaching Assistant<br>Instructor in Statistics, Experimental Methods, and Industrial/Organizational Psychology<br>Department of Psychology<br>Old Dominion University<br>Norfolk, VA |
| 1979-1980 | Graduate Research Assistant<br>Department of Psychology<br>Old Dominion University<br>Norfolk, VA |
| 1978-1979 | Graduate Teaching Assistant<br>Experimental Methods Laboratory Instructor<br>Department of Psychology<br>Old Dominion University<br>Norfolk, VA |
| 1969-1976 | Cryptologic Technician<br>Naval Security Group<br>U. S. Navy |

Honors and Awards:

| | |
|---|---|
| 1980-1981 | Awarded Doctoral Fellowship<br>Old Dominion University |
| 1979 | Named Outstanding Graduate (M.S.) Student in Psychology, Old Dominion University |

1978 Graduated Summa Cum Laude
Named Outstanding Senior in Psychology,
Old Dominion University

1977 Elected to Psi Chi, National Psychology Honor Society

Professional Associations:

Human Factors Society
Southeastern Psychological Association
Tidewater Human Factors Group
Wiregrass Human Factors Group

Publications and Presentations:

Derlega, V. J., McAnulty, D. M., Strout, S., & Reavis, C. A. (1980). Pygmalion effects among blacks: When and how expectancies occur. Journal of Applied Social Psychology, 10, 260-271.

Jones, D. H. & McAnulty, D. M. (1984). An examination of ability requirements for various rotary wing missions. Proceedings of the 28th annual meeting of the Human Factors Society. San Antonio, TX.

Jones, D. H., McAnulty, D. M., Shipley, B. D., & Sanders. M. G. (1984). An evaluation of the mission track assignment battery as a classification system for Army aviators. Phase I report: An examination of ability requirements for various rotary wing missions (Technical Report ASI 479-052-84). Fort Rucker, AL: Anacapa Sciences.

McAnulty, D. M. (March, 1981). Figural identity effects on the perception of apparent movement produced by a bistable multielement motion display. Paper presented at the annual meeting of the Southeastern Psychological Association.

McAnulty, D. M. (May, 1981). Age, S-R compatibility and stimulus uncertainty effects on information processing. Paper presented at the annual meeting of the Virginia Academy of Science.

McAnulty, D. M. (March, 1982). The age decrement in reaction time: Implications for central information processing mechanisms. Paper presented at the annual meeting of the Southeastern Psychological Association.

McAnulty, D. M. (May, 1984). Evaluation of the Army Revised Flight Aptitude Selection Test (RFAST) and the development of a future FAST. Paper presented at the semiannual meeting of the Human Factors in Aviation Screening and Performance Prediction Sub Technical Group (STAG), Fort Rucker, AL.

McAnulty, D. M. (October, 1984). Army aviator selection and classification research update. Paper presented at the semiannual meeting of the Human Factors in Aviation Screening and Performance Prediction Sub Technical Group (STAG), Brooks Air Force Base, TX.

McAnulty, D. M. (1985). Evaluation of a Flight Surgeon Course syllabus change (Research Report ASI 479-065-84). Fort Rucker, AL: Anacapa Sciences.

McAnulty, D. M. & Bierbaum, C. R. (1986). Evaluation of the AH-64 Target Acquisition Designation Sight Selected Task Trainer as a potential selection instrument (Technical Memorandum ASI 677-103-86). Fort Rucker, AL: Anacapa Sciences.

McAnulty, D. M., Cruser, L. G., & Hedge, J. W. (April, 1979). The role of polydipsia in the temporal discrimination by rats of fixed interval schedules of reinforcement. Paper presented at the annual meeting of the Southern Society for Philosophy and Psychology.

McAnulty, D. M. & Jones, D. H. (1984). An evaluation of aviator training ability requirements scale ratings. Proceedings of the 28th annual meeting of the Human Factors Society. San Antonio, TX.

McAnulty, D. M., Jones, D. H., Cohen, R. J., & Lockwood, R. E. (1984). Identification of the abilities required for effective helicopter training performance (Technical Report ASI 479-046-84). Fort Rucker, AL: Anacapa Sciences.