AFHRL-TP-89-73

2

AIR FORCE

HUMAN RESOURCES

PROTOTYPES OF COGNITIVE MEASURES FOR AIR FORCE OFFICERS: TEST DEVELOPMENT AND ITEM BANKING

Frances R. Berger
Willa B. Gupta
Raymond M. Berger

Psychometrics Inc.
13245 Riverside Drive
Sherman Oaks, California 91423-2172

Jacobina Skinner

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

May 1990

Final Technical Paper for Period September 1987 - November 1989

Approved for public release; distribution is unlimited.

LABORATORY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

Daniel L. Leighton, Colonel, USAF
Chief, Manpower and Personnel Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
May 1990

3. REPORT TYPE AND DATES COVERED
Final - September 1987 to November 1989

4. TITLE AND SUBTITLE
Prototypes of Cognitive Measures for Air Force
Officers: Test Development and Item Banking

5. FUNDING NUMBERS
C - F33615-83-C-0035
PE - 62205F
PR - 7719
TA - 18
WU - 24

6. AUTHOR(S)
Frances R. Berger Raymond M. Berger
Willa B. Gupta Jacobina Skinner

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Psychometrics Inc.
13245 Riverside Drive
Sherman Oaks, California 91423-2172

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AFHRL-TP-89-73

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The Air Force Officer Qualifying Test (AFOQT), a component of the selection system for officer commissioning programs and aircrew training, undergoes frequent scrutiny for validity, currency, and security. As part of that effort, a project was undertaken to develop measures of cognitive abilities not assessed by the AFOQT. Ten new test prototypes and three existing, unpublished tests were constructed. The objectives for the new tests were to improve linkage with Air Force officer job requirements, improve selection to officer commissioning programs, and expand the AFOQT classification utility beyond that of aircrew ability measurements. Items for all 13 tests were administered to samples of basic airmen; and subsets of newly constructed items for four tests, to samples of officer candidates. Item and test characteristics were evaluated using classical test theory and Item Response Theory analyses. Results were entered into an item banking system designed to store the item text, graphics, and statistics for the new tests. Results were promising, with most tests showing moderate to high internal consistency reliability. Some tests were found to be too difficult for airmen, but supplemental data from officer candidates suggested that item difficulty and discrimination statistics would be expected to improve for higher ability examinees in the officer applicant population for which the tests were designed. Additional administrations to appropriate samples were recommended to more

14. SUBJECT TERMS
Air Force Officer Qualifying Test
aptitude tests
item banking
mental abilities testing
officer classification
officer selection
test construction

15. NUMBER OF PAGES
50

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
UL

Item 13 (Concluded):

fully evaluate the psychometric properties of the items, as well as the incremental and differential validity of the new tests relative to the AFOQT.

AFHRL Technical Paper 89-73 May 1990

PROTOTYPES OF COGNITIVE MEASURES FOR AIR FORCE OFFICERS: TEST DEVELOPMENT AND ITEM BANKING

Frances R. Berger
Willa B. Gupta
Raymond M. Berger

Psychometrics Inc.
13245 Riverside Drive
Sherman Oaks, California 91423-2172

Jacobina Skinner

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed and submitted for publication by

Lonnie D. Valentine, Jr.
Chief, Force Acquisition Branch

This publication is primarily a working paper. It is published solely to document work performed.

SUMMARY

The development of 13 prototype tests designed to predict United States Air Force officer proficiency is described in this paper. The construction of 10 new tests, and the revision and further development of three existing unpublished tests is part of the continuing effort by the Air Force Human Resources Laboratory (AFHRL) to enhance the validity and currency of the Air Force Officer Qualifying Test (AFOQT). An adjunct task was to design a secure and efficient item banking and retrieval procedure.

The contractor staff provided AFHRL with the samples of 15 proposed tests, from which AFHRL selected the 10 named below. An iterative procedure of test review by AFHRL and construction or revision by the contractor followed, until test booklets were ready to be administered. The methods of test design of the new tests can be categorized as tutorial and non-tutorial. The tutorial tests are Chart Reading, Deductive Reasoning, Flowchart Reading, Navigator Computer, and Weather Comprehension. Salient information about these areas is provided in a two-page introduction to each test, along with rules on how to apply the information. These tests might be described as "trainability" tests, whereas the non-tutorial tests may be related either to prior experience and knowledge or to basic aptitudes (e.g., spatial/perceptual). The non-tutorial tests include three that originated with AFHRL--Figure Analogies, Pre-Navigator, and Word Discrimination--and five that were newly developed--Decoding Operations, Management Decisions, Spatial Assembly, Symbol Decoding, and Text Editing.

Preliminary field tests were conducted with available samples of airmen attending Basic Military Training because it was not feasible to administer the tests to officer applicants (the intended focus of the test designs). Classical item analyses of the airmen data revealed that many items did not meet acceptability criteria. These results confirmed concerns raised early in the project that some test designs would be too difficult for airmen samples. Consequently, arrangements were made to obtain supplemental data on selected items in four tests from officer candidates attending Officer Training School and Reserve Officer Training Corps programs. Results showed that the cadet samples scored significantly higher than did the airmen, and that many more items reached the criteria for appropriate levels of difficulty and discrimination.

Ancillary information on the items was obtained from the 3-parameter logistic Item Response Theory model. These data, along with the classical item analysis results, were recorded in an item storage system developed to bank data for the new prototype tests.

The most promising of the prototypes, based on airmen results, are the Decoding Operations, Navigator Computer, Symbol Decoding, and Figure Analogies tests. However, it was evident from officer candidate results that the Deductive Reasoning, Pre-Navigator, and Weather Comprehension tests also merit further research. It was recommended that all the new tests undergo further evaluation with officer candidate groups.

Additional research was recommended to find the intercorrelations between each new test and each current AFOQT subtest in order to determine uniqueness, if any, of the new tests. Criterion-related validity studies were also recommended, with performance measures developed from job analysis of Air Force officer positions.

PREFACE

The Air Force Human Resources Laboratory (AFHRL) is designated as the primary test development agency for the Air Force Officer Qualifying Test (AFOQT) by Air Force Regulation 35-8, Air Force Military Personnel Testing System. The current research and development (R&D) effort was undertaken as part of AFHRL's responsibility to develop, revise, and conduct research in support of the AFOQT. Work was accomplished under Task 771918, Selection and Classification Technologies, which is part of a larger effort in Force Acquisition and Distribution Systems. The project was completed under Work Unit 77191824, Officer Item Pool Development, as a contractual effort (Contract F33615-83-C-0035) by Psychometrics Inc., Sherman Oaks, CA. Ms. Jacobina Skinner represented the monitoring agency as contract manager.

The authors wish to express their appreciation of the scientific and technical guidance given by Drs. William E. Alley, Lonnie D. Valentine, Jr., and Malcolm James Ree. We wish also to express our thanks to Drs. Kurt Steuck and Tom Watson, and especially to Dr. Janos B. Koplyay, for their informative review of the test prototypes and items.

The test administration support by Mr. Roy Chollman, Mr. Refugio Gonzales, and other test examiners at the AFHRL test facility at Lackland AFB, Texas, is greatly appreciated, as is the item and test analysis support by Ms. Doris Black, Mr. Bill Glasscock, and A1C Robert W. Long of the AFHRL Information Sciences Division.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES (Concluded)

LIST OF TABLES

F.,OTOTYPES OF COGNITIVE MEASURES FOR AIR FORCE OFFICERS: TEST DEVELOPMENT AND ITEM BANKING

I. INTRODUCTION

The Air Force Officer Qualifying Test (AFOQT) has been part of the selection process for officer commissioning since 1951. To assure continued AFOQT validity, currency, and security, the subtests have been revised periodically. In October 1983, the Air Force Human Resources Laboratory (AFHRL) provided for the development of an item pool for future forms of the AFOQT that would emulate Form O (the then-current form) in content, and also mandated, for experimental purposes, the construction of new item types for areas not currently assessed in the AFOQT. A secondary requirement was that the new tests measure abilities not covered in the Armed Services Vocational Aptitude Battery (ASVAB), the test used for selection and classification of enlisted personnel.[1] Development of the new content areas is part of the continual effort to upgrade AFOQT coverage of abilities required for proficiency in officer jobs.

The AFOQT is used to aid in the selection of candidates for Officer Training School (OTS) and the Reserve Officer Training Corps (ROTC), and for aircrew classification to Undergraduate Pilot Training (UPT) and Undergraduate Navigator Training (UNT). Form P, the sixteenth version of the AFOQT, was the latest to be developed (Berger, Gupta, Berger, & Skinner, 1988). Form O and Form P contain 380 items organized into 16 subtests (Table 1).

Five composites are formed of these subtests on the basis of both construct and criterion-related validity (Berger, Gupta, Berger, & Skinner, 1990). The Pilot composite, formulated to predict Undergraduate Pilot Training success, consists of the VA, MC, EM, SR, IC, BC, TR, and AI subtests. The Navigator-Technical composite consists of the AR, DI, MK, MC, EM, SR, BC, TR, RB, GS, and HF subtests, a configuration shown to be predictive of navigator training success. The Academic Aptitude composite consists of the VA, AR, RC, DI, WK, and MK subtests. The Verbal composite is formed by VA, RC, and WK; and the Quantitative composite, by AR, DI, and MK. The latter three composites are used in selecting candidates for commissioning training.

The objectives for the new tests were to improve linkage with United States Air Force (USAF) officer job requirements, improve selection to officer commissioning programs, and expand the AFOQT classification utility beyond that of aircrew ability measurements. This paper describes newly constructed items for three existing (unpublished) tests; 10 new test prototypes; details of the test construction; preliminary field testing; results of the item analysis; and item banking. Recommendations are given for further research to assess the construct and predictive validity of the new tests and the uniqueness of these tests relative to abilities already assessed in the AFOQT.

II. TEST CONTENT

The project to develop tests with content not previously covered in the AFOQT included construction of new items for three existing unpublished tests and the conceptualization and construction of 10 new content area tests. The three existing tests were Figure Analogies, Word Discrimination, and Pre-Navigator. The new prototype tests are Chart Reading, Decoding Operations, Deductive Reasoning, Flowchart Reading, Management Decisions, Navigator Computer, Symbol Decoding, Spatial

[1] ASVAB subtests are General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Numerical Operations, Coding Speed, Auto and Shop Information, Mathematical Knowledge, Mechanical Comprehension, and Electronics Information.

Table 1. **Description of Items in AFOQT Form O and P Subtests**

Subtest	No. of items	Measures of aptitude/ability/knowledge
Verbal Analogies (VA)	25	Ability to reason and recognize relationships between words.
Arithmetic Reasoning (AR)	25	Ability to understand and reason with arithmetic relationships.
Reading Comprehension (RC)	25	Ability to read and understand paragraphs.
Data Interpretation (DI)	25	Ability to interpret data from graphs and charts.
Word Knowledge (WK)	25	Ability to understand written language through use of synonyms.
Math Knowledge (MK)	25	Ability to use learned mathematical terms, formulas, and relationships.
Mechanical Comprehension (MC)	20	Mechanical knowledge and understanding of mechanical functions.
Electrical Maze (EM)	20	Spatial ability to choose a correct path through a maze.
Scale Reading (SR)	40	Ability to read scales and dials.
Instrument Comprehension (IC)	20	Ability to determine aircraft attitude from flight instruments.
Block Counting (BC)	20	Spatial ability to "see into" a three-dimensional pile of blocks.
Table Reading (TR)	40	Ability to read tables quickly and accurately.
Aviation Information (AI)	20	Knowledge of general aeronautical concepts and terminology.
Rotated Blocks (RB)	15	Spatial aptitude to visualize and manipulate objects in space.
General Science (GS)	20	Knowledge and understanding of scientific terms, concepts, principles, and instruments.
Hidden Figures (HF)	15	Perceptual and visual imagery ability to locate simple figures embedded in complex drawings.

Assembly, Text Editing, and Weather Comprehension. Table 2 indicates the ability which each test was designed to measure and suggests the occupation for which each might be appropriate. The rationales for developing these tests are given in the next section, together with general descriptions and item constructicn procedures. Appendix A presents a taxonomy for those tests amenable to content categoriza..on. Appendix B (Figures B-1 - B-13) presents one or two sample items for each test (in alphabetical order by test title). Basic item writing instructions are given in Appendix C.

Existing Tests

Figure Analogies (FA). To address the problem of high attrition among students in UNT, a recent study examined methods for improving navigator candidate selection (Shanahan & Kantor, 1986). Scores on an experimental selection battery, the Basic Navigator Battery (BNB), were found to add to the predictiveness of the Navigator-Technical composite of the AFOQT in current use. Figure Analogies was one of the subtests in the BNB. The results of the study were the basis for selecting Figure Analogies as one of the new tests for further development.

A test that measures the ability to deduce figural relationships may add to the accuracy of predicting pilot and navigator proficiency in the visual perception of topographical features. Current AFOQT spatial subtests that may relate to the same aptitude are Rotated Blocks, Hidden Figures, Block Counting, and Instrument Comprehension. The FA test (Figure B-4) requires the identification of the relationship(s) between two given figures and the perception of the same relationship(s) between a given stimuli and one of five alternate figures. The stimuli and choices vary with the complexity of the relationship. Relationships may be in terms of shape, size, dimensionality, area of embeddedness, rotation, blank/shaded/black, and presence or absence of borders or frames.

The graphic artists who constructed these items were encouraged to vary items by using as many spatial variables as were feasible on paper. Items were pretested on project personnel to discover unintended analogies that made the item ambiguous.

Pre-Navigator (PN). Pre-Navigator was the BNB subtest that best predicted the criteria of graduation/elimination, UNT Classroom Lessons Grade, UNT Simulator Lessons Grade, and UNT Flying Lessons Grade. In fact, it appears to have been a stronger predictor than any of the separate AFOQT composites. The current version of the BNB has been available for several years as an uncontrolled test and is therefore likely to have been compromised, leading Shanahan and Kantor (1986) to recommend that new versions of the BNB be developed. The construction of additional items for the PN test represents an effort in that direction.

The original PN test covered 13 different aspects of navigator activities. A common element was simple mathematics, but it was usually only a tool to show understanding of some concept. To plan the development of new items for the PN test, the original test was categorized for item topics and for mathematical operations (see Appendix A). The item topics include flight paths, fuel consumption, compass readings, Zulu time, weather, navigator logs, dial readings, radar scopes, chart reading, and plotting sheets. Many ideas for item content were found in the various procedures workbooks for UNT students (Department of the Air Force, 1983a, 1983b, 1984). As indicated by the topics above, illustrations accompanying PN items are varied, but all items require one choice from four alternative answers. Figure B-8 presents a sample PN item.

Word Discrimination (WD). Both an early, unpublished form of WD and the new 225 items constructed for the experimental form assume broad general knowledge on the part of the examinee requisite to make fine discriminations among the given words and phrases in an item. The combination of breadth of knowledge and ability to make careful verbal distinctions may be characteristic of officer positions in general and management positions in particular (Elliott, 1988; Steuck, 1987).

Table 2. **Description of Items in Prototype Tests**

Test	Hypothesized to be predictors for pilots (P), navigators (N), or managers/leaders (M)	Measures of aptitude/ability/knowledge
Chart Reading (CR)	P, N, M	Identify topographical features on maps and legends; interpret charts, maps, and graphs.
Decoding Operations (DO)	P, N, M	Decode and interpret incomplete statements rapidly.
Deductive Reasoning (DR)	P, N, M	Learn and apply certain rules of logic.
Figure Analogies (FA)	P, N	Perceive figural relationships.
Flowchart Reading (FR)	P, N, M	Use flowcharts as a tool for organizing and planning complex activities.
Management Decisions (MD)	M	Organize information in ways that will optimize management decisions.
Navigator Computer (NC)	N	Use the Navigator Computer slide rule to solve problems of speed, time, distance, and fuel consumption.
Pre-Navigator (PN)	N	Grasp conceptual and mathematical relationships in tables, diagrams, and word problems.
Spatial Assembly (SA)	P, N	Mentally select the parts that will combine to form a given whole.
Symbol Decoding (SD)	P, N, M	Employ inductive reasoning using symbols for words and words for symbols.
Text Editing (TE)	M	Recognize corrections of spelling, grammar, and syntax that would best improve the clarity and writing style of a document.
Weather Comprehension (WC)	P, N	Relate weather information to flight planning to predict conditions affecting travel outcomes.
Word Discrimination (WD)	M	Make fine discriminations (based on general knowledge) among words and phrases.

The WD items (Figure B-13) present four alternative words, names, or phrases, of which three have some element in common. The examinee is required to select the one alternative that does not belong. The distinctions are frequently associated with word meaning, but may be based on general knowledge of well-known people (fictional and historical), geography, government, grammatical property, flora and fauna, or literature.

New Tests

Chart Reading (CR). The interpretation of information regarding the topographical features of the world is essential to certain Air Force jobs. In considering the development of a Chart Reading test to measure aptitude in this area, current AFOQT subtests and the Pre-Navigator (PN) test were reviewed to see if this test type would be redundant. Data Interpretation (DI) requires analysis and interpretation of tables and graphs. The PN test includes a few items requiring chart reading, but assumes that the examinee has had prior experience or training with the charts. It was decided that a test devoted solely to chart reading would be a useful addition to the existing tests. The CR test complements the DI test in that the data to be analyzed are in chart or map form rather than in tables or graphs. CR takes one of the several aptitudes underlying navigator performance (as measured by the PN test) and examines it in depth.

The CR test is a tutorial Air Force job-related aptitude test (Figure B-1). A preliminary training sequence provides information necessary to read and interpret charts. This tutorial contains some basic physics for navigation and also explains the use of a "PLANK" (Precise Low-Altitude Navigation Key II), a clear plastic ruler scaled to match commonly used charts. On some illustrations a PLANK is drawn to appear to cover (transparently) a portion of the map. The examinees start out sharing the same information and proceed to answer questions designed to assess their ability to locate and interpret information on the given charts and maps. The questions are multiple-choice, with five alternatives. Some charts were duplicated from navigator training manuals (Department of the Air Force, 1983c, 1985), and portions of maps prepared by the Defense Mapping Agency were excerpted for item reference.

Decoding Operations (DO). Officers in management and leadership positions in the USAF should be able to convert symbols to meaningful constructs rapidly (Steuck, 1987). Coded information is used frequently in the Air Force, and a goal of the DO test is to measure the ability to decode and interpret a statement quickly. Steuck notes that the problem-solving capability being measured by this test is the ability to hold information in working memory while solving the equation (the arithmetic required is of negligible difficulty).

The DO test (Figure B-2) includes in its instructions one table giving letter-to-number conversions and another listing codes for the basic arithmetic operations (add, subtract, multiply, divide) required to complete an equation. The examinee converts the letters to numbers and deduces the missing operation. The question type is multiple-choice but has the appearance of "fill-in" questions in that the missing operation is signaled by a blank and multiple choices do not appear below each item. The choices are the same for every item (A for "plus," B for "minus," ... E for "none of the above apply") and appear only once on each test page. Every possible combination of the given codes was covered in the pool of items.

Deductive Reasoning (DR). Of 273 topics evaluated by Air Force officers (grades 01 and 02) in terms of desirability for professional education, "logical analysis for organizing ideas to support a major point," ranked 14th (Bell, 1984). Steuck (1987), in discussing the cognitive demands on management, pointed out that the ability to reason through a series of steps or derive a conclusion from a set of statements draws upon several communication and problem-solving skills. More specifically, he cited deductive reasoning as an ability required in a variety of situations facing managers. The importance of this higher-order cognitive function to effective high-level Air Force job performance made it desirable to include a special type of Deductive Reasoning test among the new content areas.

The special nature of the DR test is its tutorial approach. Diagramming rules, sentence forms, sentence diagrams, and three sample problems with explanations are given in two pages of instructions. Examinees are also provided sample problems which illustrate the question formats used in the test. One type of question (Figure B-3) presents a Venn diagram (three overlapping circles) with named variables (e.g., red flowers, roses, scented flowers) represented to a given degree (N = none; S =

variables (e.g., red flowers, roses, scented flowers) represented to a given degree (N = none; S = some) in each overlapping area. The task is to select from among five alternatives the one sentence that best represents the diagram. The concepts in this test, "categorical syllogisms," derive directly from classical logic theory, but no prior knowledge of formal logic is required. During development of the DR test, various sources of traditional logic symbology were evaluated for use in the test. Anderson's (1980) format, a method designed to prove invalidity, was rejected because the intent of the new test was to prove the validity of a conclusion by following a set of diagram rules. It was decided to modify Barker's (1980) and Copi's (1978) diagram systems.

If future validity studies show that DR measures the ability to learn and apply the rules of deductive logic, high scorers can be assumed to have an aptitude for learning, as well as deductive reasoning ability. Steuck (1987) cites J. R. Anderson (1980), who stated that the ability to make logic judgments is related in part to knowledge of appropriate problem-solving strategies and techniques, implying that deductive reasoning can be acquired through training. "...[The ability to learn]...how to reason correctly would be an important attribute for potential managers/leaders..." (Steuck, 1987).

Flowchart Reading (FR). A flowchart is a graphic representation of the various actions (dependent upon existing conditions) that can be taken to reach a goal. The ability to organize and plan activities is important to all individuals in managerial positions. Many management programs train executives to develop flowcharts to organize, plan, budget for, and manage complex activities. Air Force officers frequently need to formally define procedures so that responsibilities can be delineated, delegated, and monitored at appropriate choice or branching points. The understanding and use of flowcharts are therefore useful skills for a competent manager. In the 1960's, several committees developed standards and symbols for flowcharting (Bohl, 1971). Those elements are used in this test, but not all standard rules are followed.

The FR test (Figure B-5) consists of descriptions of situations which are each accompanied by a flowchart and four test items. Each description specifies actions to be taken and questions to be answered in order to achieve one or more goals. The flowchart diagrams the situation with decisions points and alternative paths of action filled-in at some decision points. The four decision points not filled-in with alternatives contain numbers corresponding to the four test items. Each item has five alternatives from which the examinee selects the action or decision that accurately completes a specified segment of the diagram.

Prior knowledge of flowchart reading should not appreciably affect an examinee's test score, as selecting the correct alternative is primarily dependent on one's understanding the complexities of the situation described. To the extent that the examinee understands the goals of a problem situation and the alternative strategies for reaching them, he or she will recognize which actions or decisions are missing in the flowchart.

Management Decisions (MD). Leadership and management tasks have been identified as job requirements across officer specialties (Bell, 1984) and as training needs in the curricula of the Air Force Officer Training School (OTS), Reserve Officer Training Corps (ROTC), and Squadron Officer School (SOS) training courses (Elliott, 1988). Planning is a highly important aspect of management, and prioritizing activities or other variables in terms of importance, logic, cost effectiveness, or efficiency is basic to planning and problem solving. None of the current AFOQT subtests measures this kind of ability. In that management responsibilities have been shown to be associated with officer status, measurement of this management aptitude has a place in the evaluation of candidates for officer training.

Each MD question (Figure B-6) consists of a description of a management situation or problem followed by six statements containing bits of information, planning steps, or reasons behind a given situation. The first task is to rank order the six statements in terms of their effectiveness, efficiency, logic, degree of advantage or disadvantage, plausibility, or probability. Five response choices show four statement numbers in different rank orders. The examinee selects the rank ordering that most closely matches his or her four top-ranked statements.

There may be no "perfect" organization of steps. For test construction purposes the "best" combination was determined by having six individuals with some leadership experience rank the statements. The combination selected was the one for which there was close or perfect agreement. (Items were rejected if they resulted in wide disagreement.) The four incorrect alternatives were developed such that only some of the statements ranked high by the developers appeared early in the sequence. Only four ranked numbers were used for each response option because to include all six would reduce the probability of agreement between the correct answer and an examinee's rank ordering. Broad general topics among the MD items include policy making, financial management, personnel decisions, morale management, marketing, and information collection. In the taxonomy (Appendix A), each topic has been subdivided into a number of subjects (e.g., personnel decisions, morale, discipline, productivity, training, promotions, and assignments). For flexibility in item construction, management decisions ranged from those dealing with a single event to those involving coordination and planning of a huge industry.

Navigator Computer (NC). Since the advent of inexpensive calculators and computers, many people have never used a slide rule. However, the ability to estimate and interpolate numerical data remains important in many visual tasks involving the utilization of data. The Dead Reckoning circular slide rule is a device used in USAF navigation. The desirability of measuring one's ability to use this device resulted in the development of the experimental NC test.

The NC test (Figure B-7) is a tutorial aptitude test which requires the examinee to identify a point on the slide rule and select the corresponding answer from among five alternatives. Letters are used to identify points in order to avoid confusion with numbers. Lettered points are presented in alphabetical order starting at 9 o'clock. The response options are also listed in alphabetical order. A set of alternative responses may include the same point twice, or a value listed twice at different points. These errors would arise if an incorrect scale were chosen or the wrong value were selected from the correct scale. Different scales on the ruler are consulted depending on the metric required by a question. The metrics consist of simple proportions, time-speed-distance, and fuel consumption. The following skills--mostly quantitative--are necessary to answer the test questions correctly:

1. Remembering the required outcome. Each question asks for a value that relates to a specific scale on the slide rule. Attention to the requirement guides the examinee to the appropriate scale.

2. Estimating a value. A logical estimate of a value is necessary to report one that is not 10 times greater or less than the correct value.

3. Reading very small numbers accurately. Some of the numbers appear upside down from the viewer's position.

The NC test was designed to complement the PN test; i.e., to measure an aptitude for a specific skill in which expertise is necessary for accurate navigation.

Spatial Assembly (SA). Spatial and visualization skills are needed for various technical jobs. Graphic subtests such as Hidden Figures are currently part of the AFOQT. A test that requires locating parts that would exactly fit a whole figure represents a new measure of spatial skills that might add to the predictive value of existing graphic subtests.

Each SA question (Figure B-9) presents a whole figure accompanied by nine or ten numbered smaller figures, only four of which can be assembled to form the given whole. The examinee chooses the correct assembly from among five alternatives that list differing sets of four numbers corresponding to the small figures.

Symbol Decoding (SD). Abstract reasoning is an aptitude underlying superior intelligence, a characteristic related to management ability (Steuck, 1987). It seems intuitively correct that abstract reasoning is related to many technical and professional jobs. The Arithmetic Reasoning subtest in the current AFOQT and the experimental Deductive Reasoning test discussed earlier measure abstract reasoning but in other ways. In Symbol Decoding, examinees must be able to take instances of a rule and derive the rule by linking symbols and concepts. Some real-life professional, technical, and management problems require similar inductive derivations of relationships among the components of a system in order to act on that system. The inductive reasoning component of the SD test is what distinguishes it from the other reasoning tests.

The SD questions are grouped in sets relating to three figures (symbols) and their translations which appear at the top of the set (Figure B-10). The translations--the verbal component of an SD set--all consist of three words (irrespective of articles such as a or the) which may be viewed as three columns that each contain a concept, object, or action. In Figure B-10, for example, the concept in column 1 is "color" (red or blue). In column 2, the object or concept is "mode of transportation" (boat or canoe). Column 3 represents an action (dips, skips, or slips). In combination, the words in the translations make sense and may even rhyme (e.g., the blue boat dips; the red canoe skips). A three-part symbol accompanies each translation, with each part symbolizing a word in its adjacent translation. As the words in the columns vary from one translation to the next, so do the corresponding parts of the adjacent symbols.

The questions are of two types. In one type, a symbol is the stem and the correct translation is to be selected from five alternative translations (see sample item S1). In the second, a statement is the stem (see sample item S2), and the task is to select from among five alternative symbols the symbol that represents the given statement.

Text Editing (TE). Surveys performed by the Air Force Occupational Measurement Center (Bell, 1984) and by the Air Force Institute of Technology (Fenno, 1985) identified communication tasks as being important across Air Force occupational specialties. Sufficient data exist to establish that most officer occupational specialties require written and oral communication skills (Elliott, 1988). Effective verbal communication is central to drafting and revising drafts of correspondence and reports, whether generated by the officer or by his/her staff, and is essential for effective management (Steuck, 1987). The TE test was devised to go beyond recognizing errors in spelling and punctuation to considerations of composition, clarity, and effective writing style.

In the TE test (Figure B-11), a paragraph consisting of four flawed sentences is followed by four test items. Each item corresponds to one of the sentences in the paragraph and is composed of five possible replacements for the flawed sentence. The task is to choose that replacement which improves the paragraph in terms of clarity and flow, as well as spelling, punctuation, and grammar. The TE items are hypothesized to measure the examinee's degree of mastery of the fundamentals of English necessary for effective editing of written documents and his or her aptitude for distinguishing qualitatively among writing styles.

Weather Comprehension (WC). Aviation requires a continual assessment, by both air and ground crews, of weather conditions over time and distance. Because the ability to understand and deal with weather information may be desirable in Air Force officers, whether or not their specialty is meteorology, the construction of a test to measure the aptitude for handling weather information seemed warranted.

The tutorial approach was taken in developing the WC test so that no prior special knowledge would be required on the part of the examinee. The instructional part of the test includes a weather map and definitions of weather systems and symbols. Examples of symbols are curved lines with filled triangles (cold front) and filled semicircles (warm front) protruding from one side of the curve. Forecast variables include conditions such as rain, clouds, temperature, pressure, and winds. Drawings of vertical slices of the atmosphere are also given. In the WC test, a different Horizontal Weather Depiction chart is presented for each set of 10 items. The items consist of a stem and four alternatives. The alternatives are either verbal only or both verbal and pictorial. Two sample problems are shown in Figure B-12.

Test and Item Reviews

Fifteen new content area tests were proposed to AFHRL by the contractor. A rationale, five sample items, and test directions were provided for each proposed test. AFHRL test construction specialists selected as most promising the 10 new tests described above, based on several considerations: the degree of overlap with existing AFOQT content, linkage to officer job and task requirements, and item difficulty appropriate for officer applicants.

Test review and selection was an iterative process involving exchanges between the AFHRL specialists and the contractor. For example, AFHRL specialists identified ambiguous items in the Management Decisions test to be subsequently revised by the contractor. In addition to general considerations of item writing (Appendix C), taxonomies of content (Appendix A) were developed and applied to ensure coverage of relevant subject matter and to guide decisions as to the appropriate number of items per content category. Items were previewed by the contractor staff and revised as necessary before assembling booklets for submission to AFHRL for review. Alternate approaches were often considered. For example, during the test review process the decision was reached to use Barker's (1980) and Copi's (1978) diagram system rather than Anderson's (1980) format in the Deductive Reasoning test. Revisions to all the tests were made in response to AFHRL comments before final booklets were prepared and printed for field testing.

III. TEST ADMINISTRATION AND ANALYSIS

Field Testing

Booklets. A total of 59 different test booklets were constructed for the 13 prototype tests. The numbers of booklets per test type ranged from 3 to 10 (Table 3). Numbers of items constructed for each ranged from 112 (Spatial Assembly) to 389 (Figure Analogies). A certain number of common items appeared in all the booklets for a given test so as to detect possible sample differences in score statistics (any sample of examinees took only one booklet of a particular test title). The number of common items per booklet ranged from 12 (Figure Analogies) to 20 (Decoding Operations, Flowchart Reading, Pre-Navigator, Text Editing, Weather Comprehension, and Word Discrimination). Common items in FA, PN, and WD were drawn from unpublished Air Force tests. Common items in new tests were a subset of the newly constructed items. Unique items per booklet ranged from 32 (Management Decisions and Spatial Assembly) to 45 (Word Discrimination). Forward order (five booklets) and reverse order (five booklets) items were used in DO because it was a speeded test.

Samples. Table 4 describes the samples used in the field test of prototype items. All 59 booklets comprising the 13 tests were administered to samples of at least 300 airmen. No examinee received more than one booklet for a particular test.

Because logistic and economic constraints precluded the use of a preferred sample (i.e., civilian applicants for Air Force commissions), basic airmen constituted the only practicable group on which to obtain preliminary data for evaluating the adequacy of the 2,300 new test items. The Basic Military Training program has for many years provided a large and readily accessible source of examinees for AFHRL research and development.

Table 3. Composition of Test Booklets

Test	Total number of new items	Total number of booklets	Number of new items per booklet 1-7	8+	Number of common items per booklet
Chart Reading	192	4	44	-	16
Decoding Operations	240	10	44	-	20
Deductive Reasoning	162	4	36	-	18
Figure Analogies	389	9	43	44	12
Flowchart Reading	128	3	36	-	20
Management Decisions	144	4	32	-	16
Navigator Computer	176	4	40	-	16
Pre-Navigator	168	4	42	-	20
Spatial Assembly	112	3	32	-	16
Symbol Decoding	144	3	42	-	18
Text Editing	140	3	40	-	20
Weather Comprehension	140	3	40	-	20
Word Discrimination	225	5	45	-	20
TOTAL	2360	59			

Early in the prototype test project, however, it became evident that some of the newly designed tests would be too difficult for the airmen sample. The decision was made, therefore, to obtain supplemental data on two tests -- Deductive Reasoning (DR) and Weather Comprehension (WC) -- by administering tests bookets to OTS or ROTC cadet samples. Also, during field testing, the Navigator Computer (NC) and Pre-Navigator (PN) tests proved too difficult for the airmen; so, cadet samples were obtained for those tests as well. Samples of about 200 OTS or ROTC examinees were administered two booklets of DR and one each of WC, NC, and PN. There was insufficient time to obtain cadet samples for the difficult new tests created toward the end of the project. The OTS cadets hold baccalaureate degrees and have typically completed 2 to 4 more years of formal education than the majority of airmen; ROTC members are college or university students. The OTS and ROTC samples therefore reflected more closely the education level of the target population for which the prototype tests were designed.

Administration Procedures. Multiple test administration sessions for each booklet set were required to achieve the desired sample sizes for the collection of prototype test item data. Each booklet taken by an examinee contained items from a different test; this procedure ensured that the samples for each set of a test were independent.

Table 4. Description of Samples

Test	Sets	Sample	Range of sample sizes	Range of test dates
Chart Reading	1-4	Airmen	327-352	08-89 to 09-89
Decoding Operations	1-5	Airmen	345-378	07-87 to 11-87
	6-10	Airmen	362-400	09-87 to 12-87
Deductive Reasoning[a]	2-5	Airmen	569-723	12-88 to 03-89
	2-3	OTS	196-199	03-89
Figure Analogies	1-7	Airmen	352-359	03-85 to 01-86
	8-9	Airmen	353-373	07-87 to 08-87
Flowchart Reading	1-3	Airmen	349-386	06-89 to 08-89
Management Decisions	1-4	Airmen	340-356	08-89 to 09-89
Navigator Computer	1-4	Airmen	343-400	06-88 to 09-88
	1	ROTC	218	06-88
Pre-Navigator	1-4	Airmen	335-353	06-88 to 09-88
	1	ROTC	195	07-88
Spatial Assembly	1-3	Airmen	346-369	09-89
Symbol Decoding	1-3	Airmen	341-400	12-88 to 02-89
Text Editing	1-3	Airmen	350-387	07-89 to 08-89
Weather Comprehension	1-3	Airmen	360-393	12-88 to 02-89
	1	OTS	188	07-89 to 09-89
Word Discrimination	1-5	Airmen	347-355	11-85 to 01-86

[a] Set 1 of Deductive Reasoning was pretested with a small sample of OTS cadets, and the results were used as a guide to clarify the instructions and revise some items. Because Set 1 was different from later booklets in crucial respects, it was decided not to reuse the set number.

Time limits for each power subtest were determined after the first several administrations by noting the number of minutes required for 95% of the examinees to finish that subtest. The average became the time limit for the subsequent administration of the remaining booklets of that particular subtest. For Decoding Operations, the speeded test, the time limits were established based on the number of minutes required for 5% of the examinees to complete the test.

The practices and procedures used to administer the AFOQT at operational test sites were observed as closely as possible during collection of the prototype test item data. Major features of the manual for AFOQT administration were used. Test directions were read from the booklet. Demographics and test responses were recorded on a machine-scannable answer sheet (General Answer Sheet Type C, Westinghouse Corporation, Form 093937-001 W-2300).

Analytic Techniques

The primary analysis of performance on the prototype tests at the item level was based on "true score" or classical test theory (Gulliksen, 1950; Henrysson, 1971; Koplyay, 1981). Item difficulty ($\underline{p}$) was calculated as the proportion of examinees responding correctly to the item. The biserial correlation ($\underline{r}_{bis}$) between the item score (correct or incorrect) and total test score was used as an index of the discriminative value of each item. Data on item discrimination were obtained for both keyed and nonkeyed response options, as were values of T (mean = 50, standard deviation = 10) and statistics for quintiles of the score distribution. The T value was the mean score (standardized) of examinees selecting each option. Within each quintile, calculations were made of the frequency and percentages of examinees -- both in that quintile and in the total sample -- who chose each response option. Test reliability or internal consistency reliability was computed using Coefficient Alpha.

Ancillary information on the items was obtained from the 3-parameter logistic Item Response Theory (IRT) model (Lord & Novick, 1968) using the program BILOG II (Version 2.2) (Mislevy & Bock, 1984). The item parameters, $\underline{a}$, $\underline{b}$, and $\underline{c}$, were estimated[2] and put on a common scale using the anchor (common) item method (Ree & Jensen, 1983).

IV. ITEM BANKING

An item storage system was developed to bank data for the new prototype tests. The system was modeled on an earlier item bank developed for use in computer-assisted test construction (CATC) of future AFOQT forms (Gupta, Berger, Berger, & Skinner, 1989). Both systems link text, item graphics, and item data to facilitate the locating and combining of items for future test forms.

The complete text of non-pictorial items in the prototype tests is stored on a set of floppy diskettes. Illustrations for graphic items, such as those found in the Figure Analogies and Chart Reading tests, are stored separately on a card deck. A data tape contains a variety of information for each item including codes for identifying the items, the sample tested, and the content category from the taxonomy (Appendix A), as well as statistical data. The item identification codes were designed to allow easy cross-referencing among the text, graphics, and data components of the storage system. The principal codes developed for this purpose are identifiers for Content Area, Set, Booklet, and Item Number. Table 5 shows the identification codes for the prototype tests.

The statistics computed for each item, as described earlier in the section on analytic techniques, are stored on the data tape. Statistics are recorded separately for items tested on different samples, and codes are provided to identify the sample as airmen, OTS cadets, or ROTC cadets; the size of the sample; and the dates of testing. The statistical data include results of the classical item analysis with point biserial and biserial correlations, item difficulties, and frequency and percentage distributions of examinees selecting keyed and nonkeyed response options by quintile. IRT statistics for $\underline{a}$, $\underline{b}$, and $\underline{c}$ parameters are recorded also. Appendix D presents the file layout for the data tape. This layout corresponds to the arrangement of data on the AFOQT item data tape. Readers interested in an in-depth description of the characteristics and content of the data tape are referred to Gupta et al. (1989).

[2] No priors were established for ability estimates or $\underline{b}$ parameters. Starting values for the $\underline{c}$ parameters were set at the default of the reciprocal of the number of item options. The prior standard deviation of $\underline{a}$ was set to .28. The latter procedure was recommended by R.J. Mislevy (personal communication, 7 October 1988).

Table 5. **Item Identification Codes**

Test	Content area	Sets	Booklets	Final item number
Chart Reading	CR	1-4	88031-88034	60
Decoding Operations	DO	1-5	85175-85179	64
		6-10	85182-85186	64
Deductive Reasoning	DR	2-5	88002-88005	54
Figure Analogies	FA	1-7	84120-84126	55
		8-9	85187-85188	56
Flowchart Reading	FR	1-3	88026-88028	56
Management Decisions	MD	1-4	88041-88044	48
Navigator Computer	NC	1-4	88011-88014	56
Pre-Navigator	PN	1-4	88006-88009	62
Spatial Assembly	SA	1-3	88046-88048	48
Symbol Decoding	SD	1-3	88021-88023	60
Text Editing	TE	1-3	88036-88038	60
Weather Comprehension	WC	1-3	88016-88018	60
Word Discrimination	WD	1-5	85160-85164	65

Secondary components of the item banking system are three documents intended to assist test construction specialists in using the item bank. Printed test booklets contain hard-copy text of the test directions and show the format, size, and layout of the item text and illustrations. Item statistics contained on the data tape are also provided in hard-copy form. These printouts provide supplemental information on each test such as test reliability, standard error of measurement, and score means and standard deviations. Finally, a taxonomy of content (Appendix A) identifies and describes the subject content of the tests and codes used on the data tape to represent the specific content category of individual items.

V. RESULTS

The results of classical item analyses are summarized in Table 6. Statistics reported are raw score means and standard deviations for each booklet set; standardized item difficulties (D values) and their corresponding proportion passing; numbers of easy ($p > .70$), difficult ($p < .30$), and midrange ($p = .30 - .70$) items in each set; and internal consistency reliabilities. At the test level, mean item difficulty was computed by transforming each item difficulty (p) to a standardized difficulty (D) and averaging the D values. The mean D was then converted back to the corresponding proportion (P) to obtain information on the average proportion of items passed in the test (Koplyay, 1981, p. 61).

Table 6. **Difficulty Indices and Reliability**

Experimental test		**Items per set**	**Raw Score**		**Standardized item difficulty**		**Distribution of item difficulty (p)**			**Test relia-bility**
			Mean	**SD**	**Mean D**	**Mean P**	**>.70**	**.70-.30**	**<.30**	
Chart Reading		60								
Airmen	Set 1		15.7	5.0	.39	.25	0	17	43	.57
	Set 2		14.9	4.8	.39	.24	0	18	42	.56
	Set 3		15.9	5.8	.39	.26	0	18	42	.68
	Set 4		14.2	5.0	.38	.23	0	13	47	.59
Decoding Operations		64								
Airmen	Set 1		44.7	12.4	.72	.91	63	1	0	.96
	Set 2		41.6	13.1	.71	.90	61	2	1	.96
	Set 3		43.8	12.1	.70	.90	62	2	0	.95
	Set 4		41.4	12.7	.70	.90	62	2	0	.96
	Set 5		46.1	11.6	.73	.93	63	1	0	.95
	Set 6		41.9	13.0	.71	.91	62	0	1	.96
	Set 7		43.9	12.3	.70	.89	60	4	0	.96
	Set 8		43.0	12.3	.72	.91	63	1	0	.96
	Set 9		43.5	12.1	.73	.92	61	3	0	.95
	Set 10		44.5	12.4	.73	.92	63	1	0	.96
Deductive Reasoning		54								
Airmen	Set 2		19.4	6.5	.43	.34	2	29	23	.76
	Set 3		19.2	7.3	.43	.34	3	31	20	.81
	Set 4		20.6	7.6	.45	.37	2	30	22	.83
	Set 5		20.1	7.7	.44	.36	2	32	20	.83
OTS	Set 2		30.2	10.6	.53	.57	12	38	4	.91
	Set 3		31.3	10.3	.54	.59	14	36	4	.91
Figure Analogies		55								
Airmen	Set 1		30.8	8.9	.52	.57	14	33	8	.87
	Set 2		27.9	8.2	.50	.51	9	38	8	.84
	Set 3		26.9	8.4	.50	.49	8	35	12	.85
	Set 4		25.3	8.7	.48	.46	6	36	13	.86
	Set 5		27.3	8.6	.50	.49	8	39	8	.86
	Set 6		28.7	8.9	.51	.52	9	40	6	.86
	Set 7		27.1	9.2	.50	.49	5	44	6	.87
	Set 8	56	25.3	9.1	.48	.45	6	42	8	.86
	Set 9	56	24.2	7.8	.47	.42	3	44	9	.81

Table 6. (Continued)

Experimental test		Items per set	Raw Score		Standardized item difficulty		Distribution of item difficulty (p)			Test reliability
			Mean	SD	Mean D	Mean P	>.70	.70-.30	<.30	
Flowchart Reading		56								
Airmen	Set 1		19.5	9.0	.43	.34	0	36	20	.87
	Set 2		17.4	7.2	.42	.31	0	24	32	.80
	Set 3		17.0	6.5	.41	.30	0	28	28	.76
Management Decisions		48								
Airmen	Set 1		13.0	4.0	.40	.26	0	15	33	.44
	Set 2		13.2	4.0	.40	.27	0	14	34	.45
	Set 3		14.1	4.2	.41	.29	0	14	34	.48
	Set 4		12.4	3.9	.39	.25	0	13	35	.44
Navigator Computer		56								
Airmen	Set 1		18.4	8.1	.42	.32	0	31	25	.84
	Set 2		24.5	13.8	.47	.43	0	46	10	.95
	Set 3		19.2	9.2	.42	.33	0	32	24	.88
	Set 4		19.4	9.1	.43	.33	0	33	23	.88
ROTC	Set 1		38.6	12.6	.59	.70	29	26	1	.95
Pre-Navigator		62								
Airmen	Set 1		19.1	5.9	.42	.30	0	31	31	.65
	Set 2		20.2	6.4	.42	.32	0	38	24	.69
	Set 3		19.2	5.7	.42	.30	0	33	29	.63
	Set 4		19.4	5.5	.42	.30	0	29	33	.60
ROTC	Set 1		29.8	9.0	.49	.48	9	43	10	.85
Spatial Assembly		48								
Airmen	Set 1		17.2	6.0	.44	.35	1	30	17	.73
	Set 2		17.2	6.5	.44	.35	1	28	19	.78
	Set 3		16.4	6.7	.43	.33	0	30	18	.79
Symbol Decoding		60								
Airmen	Set 1		37.6	13.4	.56	.64	19	39	2	.95
	Set 2		34.0	13.1	.53	.58	20	34	6	.94
	Set 3		37.8	13.7	.56	.64	20	40	0	.95

Table 6. (Concluded)

Experimental test		Items per set	Raw Score		Standardized item difficulty		Distribution of item difficulty (p)			Test reliability
			Mean	SD	Mean D	Mean P	>.70	.70-.30	<.30	
Text Editing		60								
Airmen	Set 1		19.4	6.6	.42	.31	0	34	26	.73
	Set 2		20.6	6.9	.43	.33	1	35	24	.75
	Set 3		17.3	5.7	.41	.28	0	25	35	.65
Weather Comprehension		60								
Airmen	Set 1		23.5	9.2	.45	.39	0	47	13	.86
	Set 2		23.6	8.2	.45	.39	0	47	13	.81
	Set 3		23.9	8.5	.46	.39	1	39	20	.83
OTS	Set 1		34.0	11.2	.53	.57	12	43	5	.91
Word Discrimination		65								
Airmen	Set 1		42.7	5.7	.58	.68	39	20	6	.67
	Set 2		34.4	6.0	.51	.53	22	28	15	.68
	Set 3		37.9	6.0	.54	.60	23	31	11	.69
	Set 4		38.2	5.6	.54	.60	26	30	9	.63
	Set 5		38.1	5.7	.54	.60	26	29	10	.65

Score Characteristics. The largest variations in mean scores for the airmen samples across sets of booklets within individual tests were observed for NC (18.4 - 24.5), FA (24.2 - 30.8), and WD (34.4 - 42.7). In distributing test items across booklets, "equal" distributions of difficulty were attempted. Further data such as demographic information would be needed to determine whether sample characteristics accounted for the differences in mean scores for NC, FA, and WD, or whether the a priori difficulty estimates were incorrect.

The PN, SA, and WC tests showed the least variation across means. The standard deviations did not appear to vary greatly among the samples for any one test. The only exception was for the airmen sample who took Set 2 of the NC test. Their mean score and standard deviation were significantly higher than those of the airmen who took the other NC sets.

The few samples of OTS and ROTC examinees achieved much higher mean scores than did the airmen on the same tests. These differences were expected because of the different educational backgrounds of the samples, as mentioned above. Figures 1 and 2 illustrate the differences between the score distributions of airmen and those of the ROTC and OTS samples on NC (Set 1) and DR (Set 2), respectively. Officer candidates, who are older on the average than ROTC cadets, tended to have much higher mean scores and wider ranges of scores.

Item Difficulty. The statistic "Mean P" is based on the standardized item difficulty computed from the different proportions within a sample that answer the various items in a given test booklet or set correctly. The data in Table 6 in the Mean P column may be interpreted as "average proportion passing the items in a set." These data show the easier tests for the airmen to be DO, (.89 - .93), WD (.53 - .68), and SD (.58 - .64). The more difficult tests for the airmen were CR,

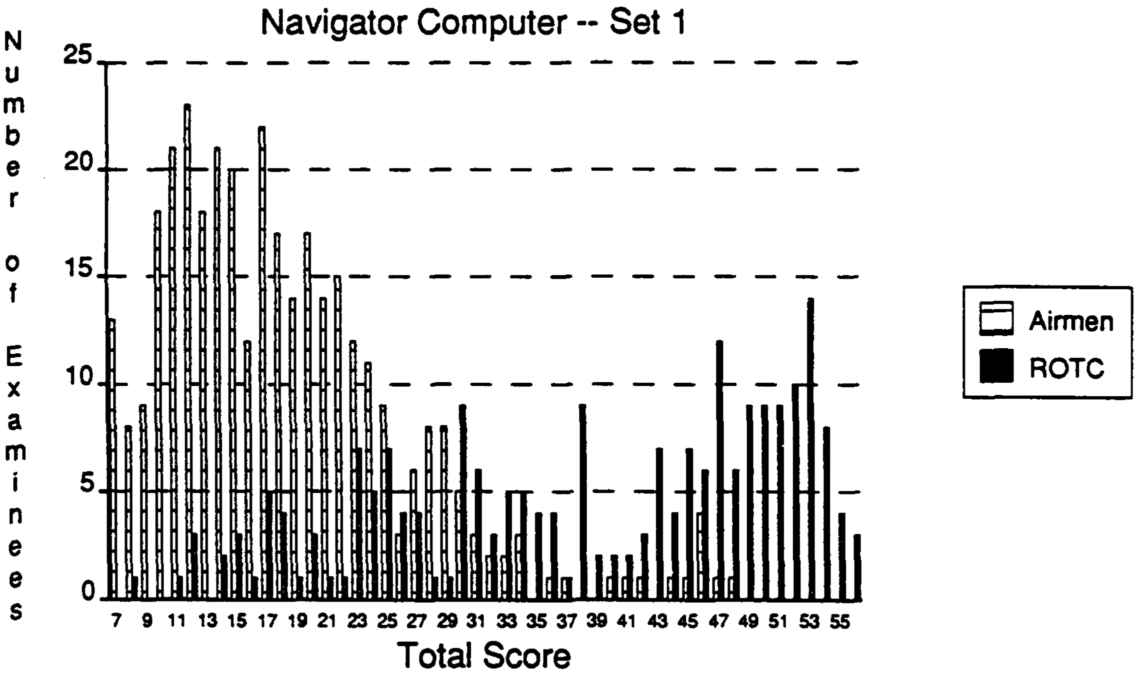


Figure 1. Score Distribution for Airmen and ROTC Cadets on Navigator Computer Set 1

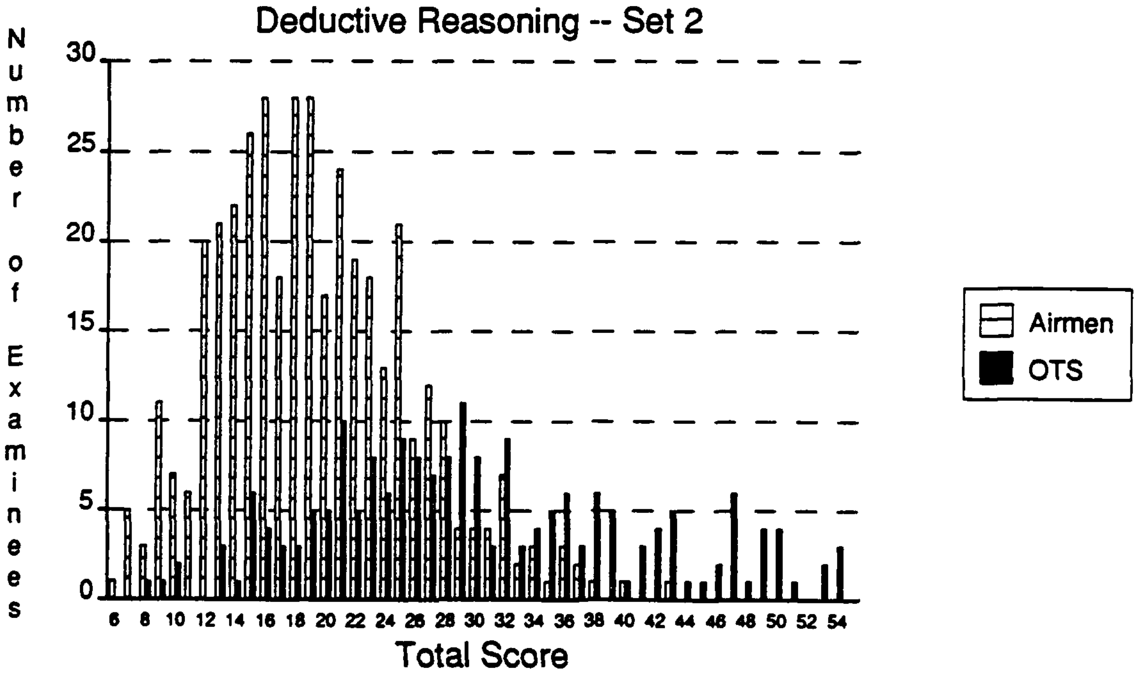


Figure 2. Score Distribution for Airmen and OTS Cadets on Deductive Reasoning Set 2

MD, TE, PN, FR, SA, DR, NC, and WC (in order of difficulty), with mean P's typically in the .30's. Mean P's were much higher for cadets than for airmen: DR (.57 and .59) and WC (.57) for the OTS samples, and NC (.70) and PN (.48) for the ROTC samples. These results, of course, are consistent with the finding of higher mean scores for the OTS and ROTC groups.

Reliability. For the airmen samples, the highest internal consistency reliabilities were those exhibited by DO (.95 - .96), SD (.94 - .95), NC (.84-.95), and FA (.81-.87). Because DO is a speeded test, its internal consistency reliabilities are probably inflated. The lowest test reliabilities were for MD (.44 - .48), CR (.56 - .68), PN (.60 - .69) and WD (.63 - .69). These reliabilities were lower than are generally considered acceptable. In the case of WD, the low reliabilities may be attributed to the diversity of subject matter in the items, though the test task appeared to be the same throughout.

The remaining tests administered to airmen (DR, FR, SA, TE, and WC) had reliabilities ranging from .65 to .87, with a mode of .83. Test reliabilities for the tests administered to the OTS or the ROTC samples were higher than those for the same test sets taken by airmen. For example, the reliability of PN Set 1 was .65 for airmen and .85 for the ROTC sample.

Item Discrimination. The item analyses provided biserial correlations (r_{bis}) for items in all the tests. The number of new items per test with r_{bis} values $\geq$.40 for the correct answer and negative correlations for the wrong alternatives appears in Table 7. (These standards are traditionally accepted for identifying items with sufficient discriminative power for inclusion in aptitude tests.) The total number of acceptable items for the 10 newly constructed tests includes a tally of common items in Set 1 and a tally of unique items from all the sets. For example, of the 128 items constructed for Flowchart Reading, 8 common items in Set 1 and 55 unique items in Sets 1, 2, and 3 met the acceptability criteria. The total number of acceptable items for FA , PN, and WD did not include the common items supplied by AFHRL.

Unique Versus Common Items. Twenty common items each were provided by AFHRL for Pre-Navigator and Word Discrimination, and 12 items for Figure Analogies. A comparison of AFHRL-supplied items (evaluating common items across all booklet sets) and newly constructed items revealed some fairly minor differences between the two sources. For the PN test, the new items were more difficult, with mean P values for new items and common items, respectively, being .44 versus .55 for ROTC examinees and .28 versus .35 for airmen. For the WD test, the percentage of new items meeting the acceptability criteria (26%) was greater than that for common items (19%). The reverse was true for both FA (52% vs. 77%) and PN (15% vs. 23%) for new items and common items, respectively.

For airmen, SD achieved the greatest proportion of acceptable items (134 of 144, or 93%), followed by NC (120 of 176, or 68%), DO (139 of 240, or 58%), and FA (198 of 384, or 52%). The smallest proportions of acceptable items for the airmen were found in MD (13 of 147, or 9%), CR (19 of 192, or 10%), PN (26 of 168, or 15%), TE (35 of 140, or 25%), and WD (59 of 225, or 26%). If the level of acceptability for items were to include biserials greater than .30, many more items would reach the acceptability level; for example, WD would have 106 acceptable items. For many items with a difficulty level above .80, few examinees selected the nonkeyed responses. Although these answers all showed negative biserials, the biserial obtained for the keyed answer was low.

It is noteworthy that the proportions of acceptable items rose considerably when the sample was OTS or ROTC. For the ROTC sample, 18 of 42 items (43%) were acceptable for PN; 52 of 56 items (93%), for NC. The OTS sample had 68 acceptable DR items out of 90 (76%) compared to the airmen sample, which had 63 good items out of 162 (39%). Also for the WC test, OTS had 42 good items out of 60 (70%) as compared to 63 out of 140 (45%) for the airmen.

Table 7. New Items Meeting Discrimination ($\underline{r}_{bis}$) Acceptability Criteria

Test	Total new test items	Number of acceptable items										Total acceptable items
		Common items set 1[a]	Unique items by set									
			1	2	3	4	5	6	7	8	9	
Chart Reading												
Airmen	192	3	2	3	8	3						19
Decoding Operations[b]												
Airmen	240	15	28	29	29	20	18					139
Deductive Reasoning												
Airmen	162	5	10	15	14	19						63
OTS	90	15	27	26								68
Figure Analogies												
Airmen	389	8	32	19	21	20	19	25	25	22	15	198
Flowchart Reading												
Airmen	128	8	23	17	15							63
Management Decisions												
Airmen	144	4	1	3	1	4						13
Navigator Computer												
Airmen	176	8	19	35	27	31						120
ROTC	56	14	38									52
Pre-Navigator												
Airmen	168	3	7	7	7	5						26
ROTC	42	13	18									18
Spatial Assembly												
Airmen	112	7	9	10	13							39
Symbol Decoding												
Airmen	144	18	39	37	40							134
Text Editing												
Airmen	140	4	15	7	9							35

Table 7. (Concluded)

		Number of acceptable items										
	Total	Common	Unique items by set									Total
	new test	items	1	2	3	4	5	6	7	8	9	acceptable
Test	items	set 1[a]										items
Weather Comprehension												
Airmen	140	10	21	16	16							63
OTS	60	14	28									42
Word Discrimination												
Airmen	225	5	10	12	13	13	11					59

Note. Acceptable items had $\underline{r}_{\text{bis}}$ ≥.40 for keyed responses and negative $\underline{r}_{\text{bis}}$ for all nonkeyed responses.

[a]Common items for FA, PN, and WD were obtained from prior unpublished versions of these tests. For these tests, the common items meeting acceptability criteria were not included in the "Total acceptable" counts.

[b]The Decoding Operations set number headings should be interpreted as "Sets 1 and 6, Sets 2 and 7, ... , and Sets 5 and 10." The same items were administered in forward order in the first set and in reverse order in the second set of each booklet pair. T'.is analysis tabulated results for items 1 - 32 in each set, thereby ensuring that any DO item meeting the acceptability criteria was counted only once.

VI. DISCUSSION

The approach to test design for several of the new prototypes was a significant departure from that used for traditional aptitude testing. The precedent for utilizing a tutorial approach was established by the use of tutorials in the AFOQT. One subtest in particular, Instrument Comprehension, included a comprehensive two-page tutorial. This two-page limit was followed for all prototype tests. The tutorial method was used in Chart Reading, Deductive Reasoning, Flowchart Reading, Navigator Computer, and Weather Comprehension. Tutorial tests such as Weather Comprehension (WC) measure skills differently than do non-tutorial tests such as Word Discrimination (WD). WC is more a "trainability" test whereas WD depends upon previously acquired knowledge. The design of tutorial tests requires sampling the subject domain in such a way that the elements taught are few enough to be learned by the examinees during a brief test administration period, yet provide sufficient coverage to be representative of proficiency in the field. There is necessarily some trial and error involved in achieving the optimal balance; so, further development of the more promising tutorial tests may be fruitful.

A rule of thumb in test development is that probably only one in three items developed (33%) will be found acceptable. By this rule, the percentages of acceptable items for Chart Reading (10%), Management Decisions (9%), Pre-Navigator (15%), Text Editing (25%), and Word Discrimination (26%) fall short, based on item-test biserial correlations for airmen samples. A noteworthy characteristic of these tests is they sample a complex domain. Two other such tests--Flowchart Reading (49%) and Weather Comprehension (45%)--also had fairly low percentages of acceptable items. The tests that fared best with airmen--Decoding Operations (58%), Figure Analogies (52%), Navigator Computer (68%), and Symbol Decoding (93%)--were more uniform in content.

The tests that sampled a more complex content domain were the more difficult ones for airmen subjects. For example, only 23% of the Chart Reading items had a difficulty value ($\underline{p}$) above .30.

The later items on the test were omitted by 30% of the sample. The responses of examinees who did attempt the items were evenly distributed across all five options, suggesting guessing. Similar results were found for Management Decisions, Spatial Assembly, Text Editing, and Flowchart Reading. Because the prototype tests were designed for higher level examinees, the item difficulty results for airmen are not unexpected.

The commentary on difficulty issues applies to most of the power tests. An exception is Word Discrimination (WD), which proved to be relatively easy for the airmen. WD had many items with difficulty values (p) greater than .70. As a result, the score ranges on the different WD booklets were small, a factor which restricts the possible magnitude of biserial correlations. The easier items tended to have negative biserials for all incorrect alternatives, but their biserials for the keyed alternatives were under .40. Based on the airmen data, WD items prepared in the current project are probably too easy to be used with the target officer applicant population. Making the WD items more difficult would be feasible, and would probably increase their discriminative ability.

Findings from the test administrations to OTS and ROTC cadets point to the need for evaluating all the prototype tests using samples whose ability and education are representative of the target officer applicant population. Test results for Deductive Reasoning, Navigator Computer, Pre-Navigator, and Weather Comprehension improved substantially when these tests were administered to cadets. The tests were shown to be more reliable and to have a greater number of acceptable items in appropriate difficulty categories than when administered to airmen. Similar improvements would be expected for other prototype tests found to be difficult for airmen (i.e., Chart Reading, Flowchart Reading, Management Decisions, Text Editing, and Spatial Assembly). In contrast, the Figure Analogies and Symbol Decoding tests, and especially the Word Discrimination test, which were shown to be quite easy for airmen, may not be sufficiently challenging for ROTC or OTS cadets.

Most of the tutorial tests and two of the non-tutorial tests, Management Decisions and Pre-Navigator, were designed to measure specialized job aptitudes. The remaining tests--Deductive Reasoning, Text Editing, Figure Analogies, Spatial Assembly, Symbol Decoding, and Word Discrimination--were based on constructs related to high-level Air Force jobs, primarily in the domains of logic, spatial visualization, and verbal ability. As discussed earlier, the tests were developed to be relevant for officer selection, especially for the jobs of pilot, navigator, and manager/leader. Additional research is needed to evaluate the utility of including the tests in the current selection system.

Data are needed on the intercorrelations among the new tests and between each new test and each current AFOQT subtest. In this regard, Ree (1989) has commented that uniqueness in prediction is one of the seminal requirements that qualifies a new predictor for a selection system. Ree has also discussed also the need to establish reliability, and recommends parallel forms correlation.

Criterion-related validity studies of the prototypes need to be conducted with suitable samples. Because the tests are intended as aptitude rather than proficiency tests, it is important to establish their predictive, rather than concurrent, validity. Performance measures identified by job analyses of officer positions, including pilot and navigator specialities and jobs with management/leadership requirements, are the desired or ultimate criteria. The data gathered for validity studies can be used to determine whether composites of the new tests alone or with selected AFOQT subtests would be better predictors of officer performance in different occupations than are the individual tests. Validation of the prototype tests, whether individual or combined, for predicting performance in specific specialities should be undertaken with appropriate samples of applicants or commissioned officers.

REFERENCES

Anderson, J.R. (1980). *Cognitive psychology and its implications*. San Francisco: Freeman and Company.

Bell, J.M. (1984). *Special report: Professional military education - officer* (AFPT 90-XXX-522). Randolph AFB, TX: USAF Occupational Measurement Center.

Barker, S.F. (1980). *The elements of logic*. New York: McGraw-Hill.

Berger, F.R., Gupta, W.B., Berger, R.M., & Skinner, J. (1990). *Air Force Officer Qualifying Test (AFOQT) Form P: Test manual* (AFHRL-TR-89-56). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Berger, F.R., Gupta, W.B., Berger, R.M., & Skinner, J. (1988). *Air Force Officer Qualifying Test (AFOQT) Form P: Test construction* (AFHRL-TR-88-30, AD-A200 678). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Bohl, M. (1971). *Flowcharting techniques*. Chicago: Science Research Associates.

Copi, I.M. (1978). *Introduction to logic* (5th ed.). New York: Macmillan.

Department of the Air Force. (1983a, April). *Basic procedures* (Study Guides/Workbooks N-V6A-F-BP-SW). Mather AFB, CA: Undergraduate Navigator Training.

Department of the Air Force. (1983b, April). *Navigation procedures* (Study Guides/Workbooks N-V6A-F-NP-SW (Vol. 1)). Mather AFB, CA: Undergraduate Navigator Training.

Department of the Air Force. (1983c, April). *Night celestial navigation* (Study Guides/Workbooks N-V6A-F-NC-SW). Mather AFB, CA: Undergraduate Navigator Training.

Department of the Air Force. (1984, December). *Airmanship* (Study Guides/Workbooks N-V6A-F-AM-SW (Vols. 1-3)). Mather AFB, CA: Undergraduate Navigator Training.

Department of the Air Force. (1985, July). *T-43 flight training manual* (Flight Training Manual N-V6A-F-FTM T-43). Mather AFB, CA: Undergraduate Navigator Training.

Department of the Air Force. (1986, July). *Weather* (Student Workshop S-V8N-C-CWX-SW). Mather AFB, CA: Undergraduate Navigator Training.

Elliott, L.R. (1988). *The identification of abilities and/or aptitudes relevant to the selection of AF officers*. Unpublished manuscript, Manpower and Personnel Division, Air Force Human Resources Laboratory, Brooks AFB, TX.

Fenno, C.R. (1985). *A profile of the communication tasks of Air Force officers* (Technical Report No. AU-AFIT-LS-1-85). Wright-Patterson AFB, OH: Air Force Institute of Technology. (NTIS No. AD-A163 476)

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons, Inc.

Gupta, W.B., Berger, F.R., Berger, R.M., & Skinner, J. (1989). *Air Force Officer Qualifying Test (AFOQT): Development of an item bank* (AFHRL-TP-39-33, AD-A216 228). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Henrysson, S. (1971). Gathering, analyzing and using data on test items. In R.L. Thorndike (Ed.), *Educational measurements* (pp. 130-159). Washington, DC: American Council on Education.

Koplyay, J.B. (1981). *Item analysis program (IAP) for achievement tests* (AFHRL-TP-81-22, AD-A107 884). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental scores*. Reading, MA: Addison-Wesley Publishing Co.

Mislevy, R.J., & Bock, R.D. (1984, June). *BILOG II: Item analysis and test scoring with binary logistic models (version 2.2)*. Mooresville, IN: Scientific Software, Inc.

Ree, M.J. (1989, November). *Validation of new predictors*. Paper presented at the meeting of the Military Testing Association, San Antonio, TX.

Ree, M.J., & Jensen, H.E. (1983). Effects of sample size on linear equating of item charateristic curve parameters. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 135-146). New York: Academic Press.

Shanahan, F.M., & Kantor, J.E. (1986). *Basic navigator battery: An experimental selection composite for undergraduate navigator training* (AFHRL-TR-86-3, AD-A168 857). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Steuck, K.W. (1987, November). *Air Force Officer Qualifying Test (AFOQT): Promising cognitive measures of leadership and management*. Paper presented at the meeting of the Texas Psychological Association, San Antonio, TX.

APPENDIX A: TAXONOMY OF TEST CONTENT

Test	Content Code	Description
Chart Reading		

Small sections of Operational Navigation Charts were reproduced for illustrations in the CR test. The chart type labels are arbitrary; they indicate different types of maps that were used.

	A1	Chart type A #1
	A2	Chart type A #2
	BC	Chart type B (common)
	CC	Chart type C (common)
	C1	Chart type C #1
	C2	Chart type C #2
	C3	Chart type C #3
	C4	Chart type C #4
	DC	Chart type D with PLANK (common)
	D1	Chart type D with PLANK # 1
	D2	Chart type D with PLANK # 2
	GF	Graph of fuel mileage
	GP	Graph of Pressure Attitude and Temperature Deviation (common)
		GT Graph of Time to Climb
	P1	Plot 1
	P2	Plot 2
	WM	World map -- magnetic variation
	WT	World map -- time zones
Decoding Operations		

Keyed response is the category.

	A	Add
	B	Subtract
	C	Multiply
	D	Divide
	E	None of the above apply
Deductive Reasoning	1	Stem is diagram, request statement
	2	Stem is two statements, request conclusion
	3	Stem is two statements, request diagram
Flowchart Reading	D	Decision points or questions
	A	Actions

APPENDIX A: (Continued)

Test	Content Code	Description
Management Decisions		
	D	Defining priorities; recognizing importance
	A	Problem analysis
	P	Policy making
	F	Financial management
	T	Temporal ordering
	L	Logical ordering
	E	Employment decisions
	M	Morale management
	S	Marketing
	I	Information collecting
Navigator Computer		
	Each item is coded for three dimensions.	
		Computer dials shown
	1	One computer setting
	2	Two computer settings
		Solutions required
	P	Proportion
	T	Time-speed-distance
	F	Fuel consumption
		Response options
	P	Points only presented
	A	Answers and points presented
Pre-Navigator		
	Each item is coded for two dimensions.	
		Topic
	01	Flight path
	02	Flight path with time
	03	Flight path with wind
	04	Fuel
	05	Degrees
	06	Compass
	07	Zulu time
	08	Weather
	09	Navigator log
	10	Dial reading
	11	Radar scope
	12	Plotting sheets
	13	Chart reading (Distance/Climb/Altitude)
	14	Non-specific

APPENDIX A: (Concluded)

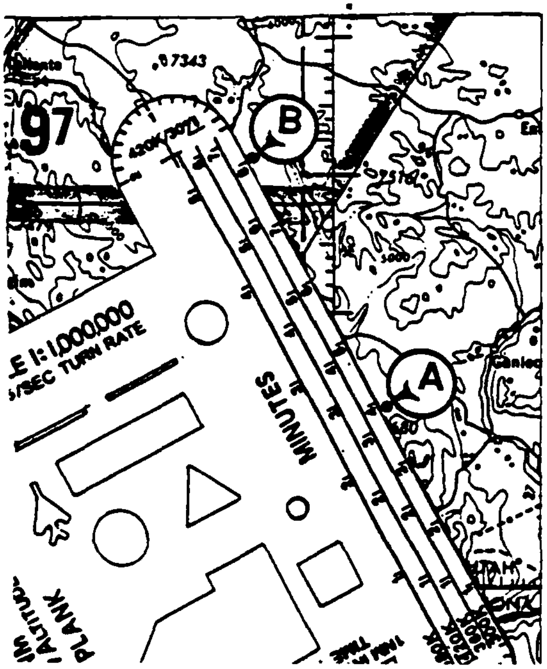
Test	Content Code*	Description
Pre-Navigator (concluded)		
		Mathematical Operation
	AL	Algebraic manipulation
	AS	Arithmetic single function
	AM	Arithmetic multiple function
	PP	Proportion or percent
	TT	Trigonometry triangles
	TA	Trigonometry arithmetic
	GR	Tables/graphics -- reference only
	GA	Tables/graphics -- arithmetic
	IN	Interpolating
	SC	Scaling
	DC	Degrees in circle
Symbol Decoding		
	1	Stem is symbol, request translation
	2	Stem is translation, request symbol
Weather Comprehension		

Items are grouped in sets of 10. (Each set refers to a different map.) The content codes refer to the item position in any set of 10. The item descriptions are the same for a given position in all sets. For example, items 1, 11, 21, 31, 41, and 51 will all have the code 1 description.

	1	Weather forecasts: give city, ask for forecast
	2	Vertical slices: 1 picture and 4 cities
	3	Weather forecasts: give forecast, ask for city
	4	Vertical slices: 1 picture and 4 cities
	5	Vertical slices: 4 pictures and 1 city
	6	Wind: while flying -- destination/wind
	7	Wind: city -- city/wind
	8	Pressure: cities/aircraft/isobars
	9	Flying: safety/fronts/combination of conditions
	0	Flying: safety/fronts/combination of conditions

Note. Tests not included in the taxonomy are those not amenable to categorization. These are Figure Analogies, Spatial Assembly, Text Editing, and Word Discrimination.

APPENDIX B: SAMPLE ITEMS FOR THE TESTS



S1. Using the PLANK, determine approximately how long it will take an aircraft, traveling at 420 knots, to get from point B to point A.

S1-A 2 minutes 30 seconds

S1-B 2 minutes 45 seconds

➤ S1-C 2 minutes 55 seconds

S1-D 3 minutes 25 seconds

S1-E 4 minutes 10 seconds

Figure B-1. Chart Reading.

CODES

AF = 10	GE = 6	NB = 5	TH = 4
BZ = 4	HC = 3	OK = 2	VG = 12
CN = 11	JY = 7	PL = 1	WS = 5
DQ = 7	KR = 1	QA = 6	XM = 9
EV = 9	LW = 2	RO = 8	YD = 8
FJ = 12	MX = 10	SP = 3	ZT = 11

OPERATIONS

A. plus

B. minus

C. multiplied by

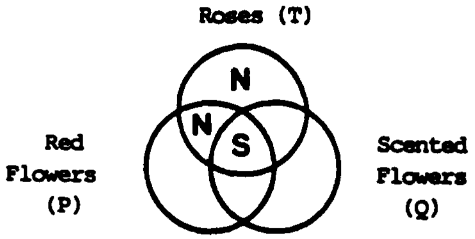
D. divided by

E. none of the above apply

S1. RO ____ HC equals ZT. Keyed Answer = A

Figure B-2. Decoding Operations.

Rules: Four possible statements relate two categories such as P and Q: all P are Q; no P are Q; some P are Q; some P are not Q. The third category is labeled T. In a diagram each is represented by a circle overlapping each of the others. S is put where there is some element, N where there is no element and blank when no information is given. The three categories are described by: statement 1 which relates Q to T or T to Q, statement 2 which relates P to T or T to P, and the conclusion which relates P to Q.



S1-A All scented flowers are roses.
S1-B Some roses are not scented flowers.
S1-C No red flowers are roses.
S1-D Some red flowers are scented flowers.
➤ S1-E Some red flowers are roses.

Figure B-3. Deductive Reasoning.

Select that figure from the alternatives (A through E) which bears the same relationship to the single figure (Z) that the two original figures (X and Y) bear to each other.

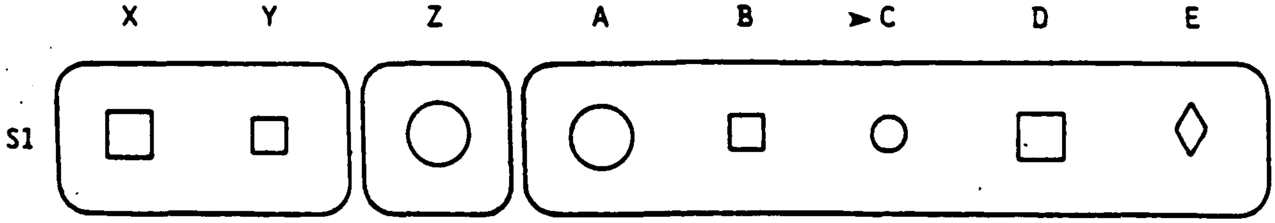
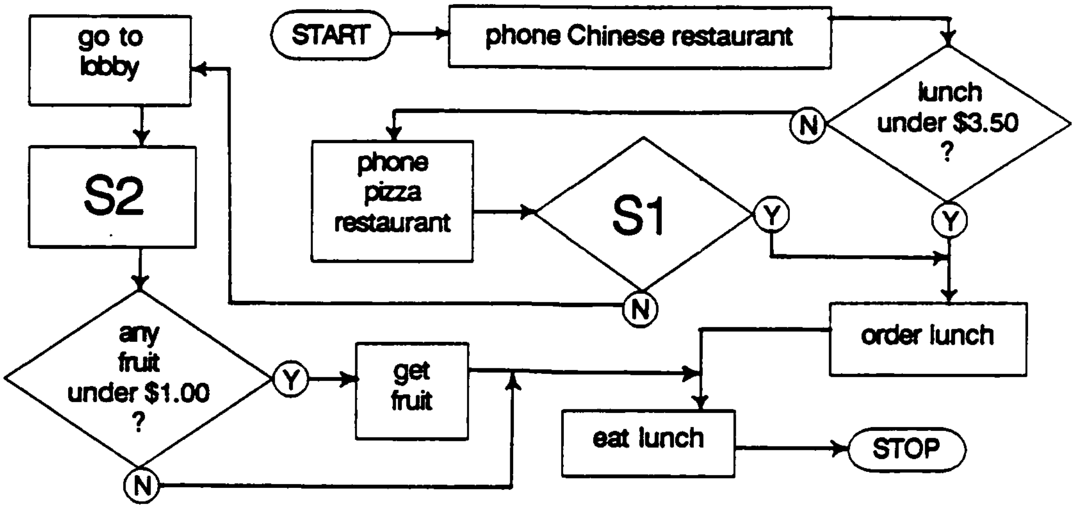


Figure B-4. Figure Analogies.

I only have $3.50 cash for lunch today. Both the Chinese restaurant and the pizza restaurant deliver, but I wonder if they have anything I can afford? I know the machine in the lobby has sandwiches for $2.00 to $3.00 as well as other items for sale.



S1-A order lunch
S1-B order a pizza
S1-C lunch for $3.00?
➤ S1-D lunch under $3.50?
S1-E lunch under $4.00?

S2-A get a cheese sandwich
S2-B get a sandwich for $2.00
➤ S2-C get a sandwich for $2.50
S2-D get an apple and banana
S2-E get fruit and potato chips

Figure B-5. **Flowchart Reading.**

S1. You are the manager of a 35-person typing and clerical department that produces printed communications for important internal use in your company (memos, forms, sales reports, newsletters, and budgets). The production schedule is very heavy this week--forms have run out and reports are needed urgently for meetings. All three of your photocopy clerks (who earn $3.75 an hour) are out with the flu. Order the actions below for efficiency and economy in solving this problem.

1. Hire three temporary photocopy personnel (who each earns $12.00 an hour) through an agency.
2. Pitch in with your assistant managers and do the work yourselves.
3. Divert three typists from their work to do the photocopying.
4. Have the neighborhood print shop, who does smudgy work, but whose charges are very low, do the work.
5. Have the remaining 32 personnel in your department contribute 15 minutes a day each to photocopying during the crises.
6. Delay production until the clerks return to work, because printing is for within-company use.

S1-A	3, 4, 5, 1
S1-B	4, 5, 1, 2
➤ **S1-C**	4, 5, 1, 3
S1-D	5, 4, 3, 6
S1-E	5, 4, 2, 6

Directions:

1. Rank all six statements.
2. Select the closest match to the four you ranked highest.

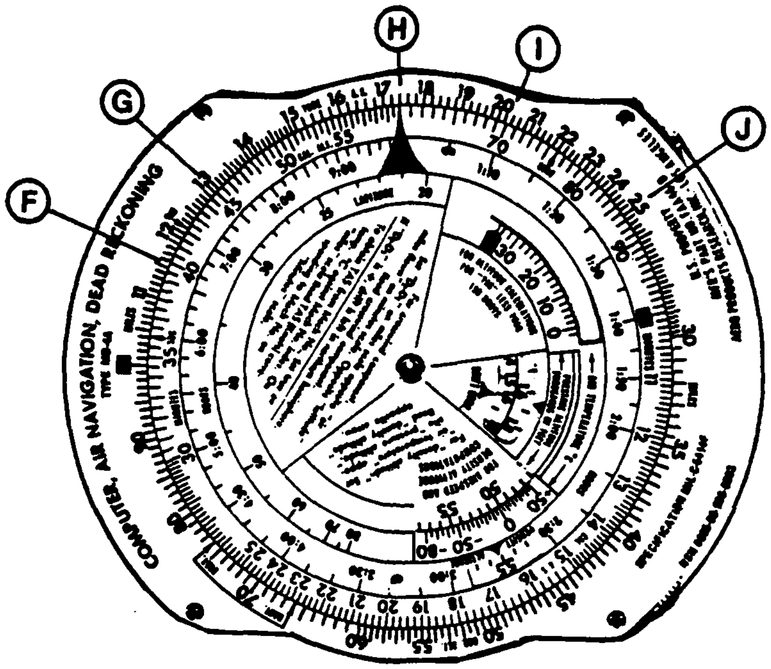
Explanation: "C," the correct answer, leads off with Action 4, because it satisfies both economy and efficiency. Professional looking quality can be sacrificed, since the distribution is for in-house purposes only. No personnel have to be taken off their jobs during a busy week.

Action 5 is ranked second because 15 minutes a day should have an almost negligible effect on production for even the higher-paid personnel in this department.

Action 1 is ranked third. It is less economical than Action 5 but less disruptive to production than Action 3.

Action 3 is ranked after Action 1, because diverting three typists full time would effect production significantly. However, it is preferable to Action 2, which would tie up top management with a clerical function, and to Action 6, which would cause an intolerable halt to communication.

Figure B-6. **Management Decisions.**



S1. A truck traveling over rough terrain covers 13 miles in 45 minutes. How far will the truck travel in 70 minutes?

S1-A	Point F
S1-B	Point G
S1-C	Point H
➤ S1-D	Point I
S1-E	Point J

S2. How far will an aircraft, averaging 174K, travel in 40 minutes?

➤ S2-A	Point F - 116 miles
S2-B	Point G - 130 miles
S2-C	Point I - 203 miles
S2-D	Point J - 250 miles
S2-E	Point J - 87 miles

Figure B-7. **Navigator Computer.**

S1. Zulu time is used by all navigators to ensure universal standardization. Seattle time is Zulu time minus 8 hours; New York time is Zulu time minus 5 hours.

A flight departs from Seattle at 1800 Zulu time and arrives in New York 5 hours later. Following a 1 1/2 hour layover, the flight returns to Seattle in 5 1/4 hours. At what local time in Seattle does this flight land?

S1-A	0845
S1-B	1345
S1-C	1645
➤ S1-D	2145

Question S1 is a sample of one type of question presented in this test.

Figure B-8. **Pre-Navigator.**

The sample (S1) shows a figure and 9 numbered pieces that might fit into this frame. Select the correct 4 piece combination.

S1.

S1-A	2, 3, 7, 8
S1-B	1, 2, 3, 4
S1-C	1, 2, 3, 6
S1-D	2, 3, 5, 9
➤ S1-E	2, 3, 4, 6

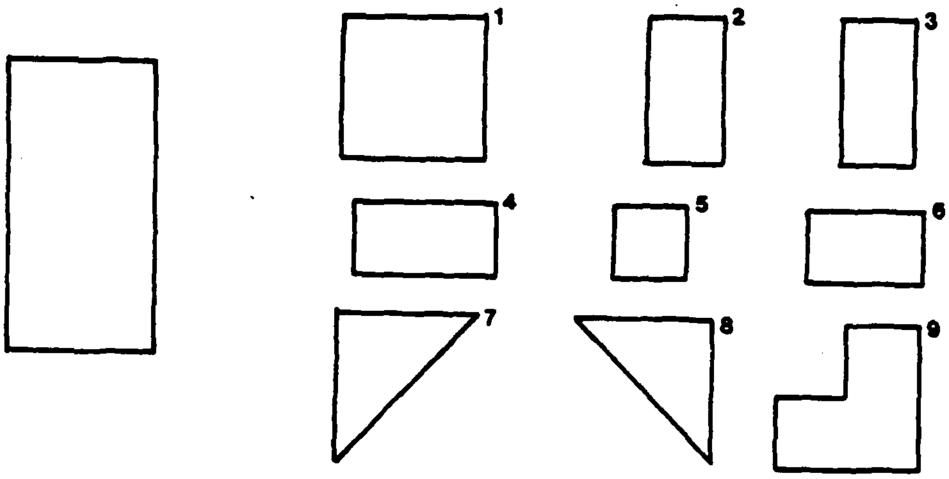
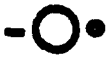
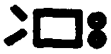
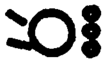
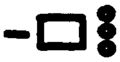


Figure B-9. Spatial Assembly.

S1 and S2 are based on the following information:

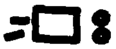
Symbol	English Translation
	The blue boat dips
	The red canoe skips
	The red boat slips

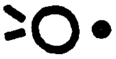
S1. Translate this symbol:

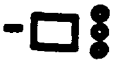


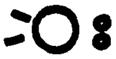
S1-A The blue canoe dips
S1-B The red boat skips
➤ S1-C The blue canoe slips
S1-D The red boat dips
S1-E The red canoe skips

S2. Which symbol represents:
The red boat skips

S2-A 

S2-B 

S2-C 

➤ S2-D 

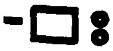
S2-E 

Figure B-10. Symbol Decoding.

S1. As a strong proponant of soil consurvation, Miss Smith's new job title will be a big help in promoting this cause.

S1-A Miss Smith, a strong proponant of soil consurvation, has a new job title that will help to promote this cause.

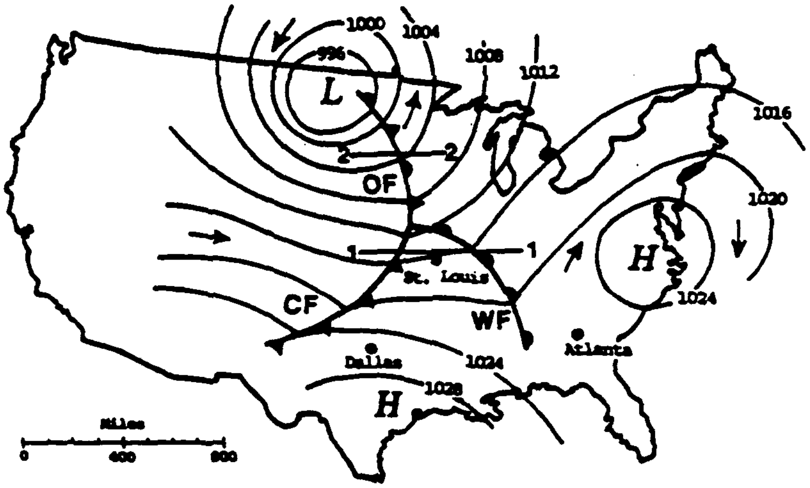
S1-B Miss Smith's new job title, being a strong proponant of soil conservation, will help to promote this cause.

➤ S1-C Her new job title will enable Miss Smith, a strong proponent of soil conservation, to promote this cause.

S1-D With her new job title Miss Smith who strongly supports soil conservation can promote this cause.

S1-E Her new job title in hand, Smith can go forward to promote soil conservation, which is strongly supported by her.

Figure B-11. **Text Editing.**



S1. The HWD chart represents the weather at 6:00 a.m. in Atlanta. The weather service in Atlanta forecasts

S1-A clear skies all day.

S1-B heavy rains starting at noon.

S1-C light rain starting at noon.

➤ S1-D morning light rain with skies starting to clear in the evening.

S2. A jet plane flying from Atlanta to Dallas would probably be flying

S2-A through a cold front to a warm front.

S2-B through a warm front to a cold front.

➤ S2-C into head winds predominantly.

S2-D with tail winds predominantly.

Figure B-12. **Weather Comprehension.**

One of these words is in some way different from the other three words.

S1-A California

S1-B Florida

➤ S1-C Chicago

S1-D Tennessee

Figure B-13. **Word Discrimination.**

APPENDIX C: INSTRUCTIONS FOR ITEM WRITERS

1. Follow the format of the sample items. All items are multiple-choice with five alternatives (except for Pre-Navigator, Weather Comprehension, and Word Discrimination, which have four alternatives).

2. Order item alternatives in ascending or descending order of length.

3. Order numeric alternatives in ascending or descending order.

4. Express items clearly in language appropriate for a high school reading level.

5. Avoid unnecessary repetition in the alternatives by including as much of the relevant information as possible in the item stem. For example, "The best way to estimate cost is through the use of:" is preferred to ending the stem with "is" and preceding all the alternatives with "through the use of."

6. Use the same terms and definitions consistently across items. For example, use either an abbreviation or a whole word consistently.

7. Avoid absolutes such as "always" and "never."

8. Ensure that all alternatives have the same grammatical structure. For example, use the same tense throughout a question's alternatives; use one voice (active or passive) so that unusual structure doesn't give away key.

9. Avoid ambiguous or vague terms. For example, a specific time reference (hourly, monthly) is preferable to "frequently."

10. Avoid colloquialisms; use standard English.

11. Avoid inclusion of nonfunctional words or unnecessary detail to keep items as short and concise as possible.

12. Do not use "none of the above" as an item alternative.

13. If an item stem contains factual information, ensure that the information is accurate.

APPENDIX C: (Concluded)

14. Write items such that there is only one correct answer possible among the alternatives. That is, provide a reasonable basis for response selection.

15. Insofar as possible, write items that reflect aspects of the work for which examinees are being tested. That is, the relevant reference is military jobs in the Air Force.

16. Avoid non-relevant clues to the correct response. Examples: making the correct alternative stand out by having it quite different from the other four in grammar, length, vocabulary, etc.; making it obvious by having the wrong alternatives appear to be silly and therefore transparently wrong.

17. Avoid sources of difficulty (e.g., unfamiliar language or symbols) that are not directly related to the content area tested.

18. Vary the difficulty of items. The number of items you are asked to write may not be enough to cover all levels of difficulty in the right proportions, and it is next to impossible to know the exact level of difficulty of an item prior to its testing. However, insofar as possible, write approximately one-third of your items to be low in difficulty, one-third to be of medium difficulty, and one-third to be of high difficulty.

19. Make the item stem informative to the point that the question can be understood before reading the response alternatives.

20. Do not write items that contain controversial material regarding sensitive issues such as morality, religion, politics, ethnicity, or regionality.

APPENDIX D: FILE LAYOUT FOR PROTOTYPE ITEM DATA TAPE

Record	Columns	Format	Variable description
0	1-2	A2	Content Area Identifier
	3-4	A2/I2	Booklet Set Number
	5	A1	Subject Type: A=Airmen, C=OTS cadets, R=ROTC
	6-7	A2/I2	Item Number 8 Al/Il Record Number "0"
	9-13	AS/IS	Booklet Number 14 Al Keyed Response: A-E or A-D
	15	A1	Speeded: F=Forward, B=Backward, Blank=Power
	16	A1	Item Type: C=Common, E=Unique each test
	17	A1	Subject Type: A=Airmen, COTS, R=ROTC
	18-20	I3	Sample Size (for that testing)
	21-28	A8/I8	Testing Date: "MoYr" for first and last month of testing.
	29		Blank
	30-33	A4	Content Category Identifier[a]
	34-37	A4	Illustration Identifier "X"
	38-41	A4	Duplicate Item Identifier (FA ony)
	42-44	A3	Line Length only)
	45		Blank
	46-72	3(F9.5)	IRT: a, b and c (overflow values = 9.99999)
	73		Blank
1	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "1"
	9-14	F6.4	Item Difficulty (Raw)
			Classical Item Analysis, Response A
	15-20	F6.4	Biserial Correlation (-1.0 set to --9999)
	21-23	I3	Percent giving that response
	24-25	12	T value that response
			Repeat 3 variables Responses B - E
	26-36		Response B
	37-47		Response C
	48-58		Response D
	59-69		Response E
	70-73		Blank
2	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "2"
	9-14	-	Blank Classical Item Analysis, Response A
	15-20	F6.4	Point Biserial Correlation
	21-23	I3	Number of Examinees giving that response
	24-25	-	Blank Repeat variables, Responses B - E
	26-36		Response B
	37-47		Response C
	48-58		Response D
	59-69		Response E

APPENDIX D: (Continued)

Record	Columns	Format	Variable description
3	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "3"
	9-26		Blank
	27-34		Test Score Range for Quintile 1
	27		Blank
	28-29	I2	Lowest Score that Quintile
	30-31	I2	Highest Score that Quintile
	32-34	I3	Number of Examinees that Quintile
	35-42		Test Score Range for Quintile 2
	43-50		Test Score Range for Quintile 3
	51-58		Test Score Range for Quintile 4
	59-66		Test Score Range for Quintile 5
	67-73		Blank
4	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "4"
			Quintile Statistics for Response A
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examlnees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5
5	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "5"
			Quintile Statistics for Response B
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examinees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5
6	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "6"
			Quintile Statistics for Response C
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examinees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5

APPENDIX D: (Concluded)

Record	Columns	Format	Variable description
7	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "7"
			Quintile Statistics for Response D
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examlnees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5
8	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "8"
			Quintile Statistics for Response E[b]
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examinees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5
9	1-7	A7	ID: (Refer to Record "0")
	8	A1/I1	Record Number "9"
			Quintile Statistics for Blanks
	9-21		Statistics for Quintile 1
	9-11	I3	Number of Examinees
	12-16	F5.1	% Examinees that Quintile
	17-21	F5.1	% of Total Examinees
	22-34		Statistics for Quintile 2
	35-47		Statistics for Quintile 3
	48-60		Statistics for Quintile 4
	61-73		Statistics for Quintile 5

[a] See Appendix A for valid codes.

[b] When there is no E response, a "0" appears in the rows or columns allotted to the E response position.