# Implementation of a Model Adaptive Testing System at an Armed Forces Entrance and Examination Station

Malcolm James Ree
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

In a world of increasing technical complexity and diminishing resources, it is the task of the military recruiting agencies to obtain the most highly qualified candidates for technical training. Traditionally, paper-and-pencil multiple-aptitude test batteries have been administered to applicants of a wide range of abilities. These tests have been peaked to be most discriminating over a relatively narrow range, because limited time precluded administration of enough items to gain maximal test information over the broad range of an ability. Selection and classification decisions must be made, however, which require discriminations at the 80th percentile. At this level, only limited information is available from a typical peaked test.

Adaptive testing, particularly computer-driven adaptive testing, promises (1) to enable the gathering of test information (Lord & Novick, 1968, Chap. 20) at all levels of ability with equal precision and (2) to increase the predictive validity of military accession testing. Furthermore, adaptive testing promises to reduce the time required to obtain ability estimates for applicants; possibly, by making accession a one-day process, it may reduce overall costs.

The model adaptive testing system was implemented in an Armed Forces Entrance and Examination Station (AFEES) in order to study its feasibility for use in a military selection setting. At the AFEES, the testing system must (1) be operated by individuals without any special training in computer hardware or software, (2) perform when needed, (3) be operational for the entire workday, (4) accommodate applicants for military service from very low ability to very high ability, (5) not intimidate or frighten the applicants or the test administrators, and (6) provide valid and reliable measurement.

Prior to the implementation of an adaptive testing system, a number of decisions--both technical and administrative--must be made. The technical questions include: (1) who are the subjects; (2) what ability areas are to be tested; (3) what items and item statistics are available; (4) which scoring method will be used; (5) which item selection technique will be used; (6) what media for question presentation will be used; and (7) how will pictorial items

be presented. There are also many administrative questions. How can the operation be simplified so that low ability, careless, or inattentive examinees do not cause an abnormal termination of testing? What impact will the demonstration have on day-to-day AFEES operations?

## Implementation

### The Setting

The San Antonio, Texas, AFEES was chosen as the test site because it was close to the development center at the Air Force Human Resources Laboratory (AFHRL). This proximity afforded considerable opportunity for monitoring the progress of the adaptive testing system.

The subjects for this demonstration were applicants for military enlistment, and their abilities covered a very broad range of aptitudes. They were tested in three aptitude areas which comprise the Armed Forces Qualification Test (AFQT): Word Knowledge (WK), Arithmetic Reasoning (AR), and Space Perception (SP). The AFQT is used for initial qualification for military service. Other aptitude areas are usually measured only if an acceptable score on the AFQT is achieved; these subjects were tested while awaiting the results of the AFQT.

### Test Items

The items used for this model adaptive testing system were culled from existing historic item files at the AFHRL. Only item difficulty (p) and item discrimination (phi) indices were available. Items were selected to represent a generally rectangular distribution of difficulties from about .2 to about .8, with the highest available discrimination index at each difficulty level. These items were then assembled into booklets for administration to Air Force basic recruits in order to estimate the latent trait parameters $a$, $b$, and $c$ (Lord & Novick, 1968) for later phases of this demonstration. Initially, the classical item indices were transformed via approximations and used to calculate the latent trait parameters useful for the project. Although these estimates varied somewhat from more exact estimates obtained from the new response data, they did permit a reasonable starting point from which to demonstrate the feasibility of adaptive testing for military service applicants. As soon as a satisfactory sample has been collected and the parameters estimated, the approximated parameters will be replaced by the new, more exact parameters.

### Computer System

A medium size computer (IBM 360/65) was available for the demonstration in a time-shared mode. The APL programming language (Gilman & Rose, 1970) was selected because it is interactive and has extremely powerful operators. In addition, experience has shown that APL leads to fast program development. It is also fast in execution and is particularly suited for handling vectors and matrices.

A combination of Bayesian item scoring and ability estimation (Owen, 1969) and selection of items by maximum information (Lord & Novick, 1968, Eq. 20.4.1)

was selected for ease of programming and low computer core utilization. Two criteria for the termination of item administration are (1) the reduction of the posterior variance of the ability estimate to a low value (<.0625) and/or (2) the subjects having taken 20 items. This procedure is also advantageous because it does not require a structured item pool as would a stratified adaptive test; thus implementation of the testing system is made easier.

A modified Tektronix model 4006-1 Cathode Ray Tube (CRT) terminal was used for the demonstration. Both a viewing hood to reduce glare and a keyboard cover to prohibit pushing inappropriate keys were fabricated. This terminal used the Tektronix Graphics Package, APLgraph 2, and was run at 1200 baud in half-duplex mode.

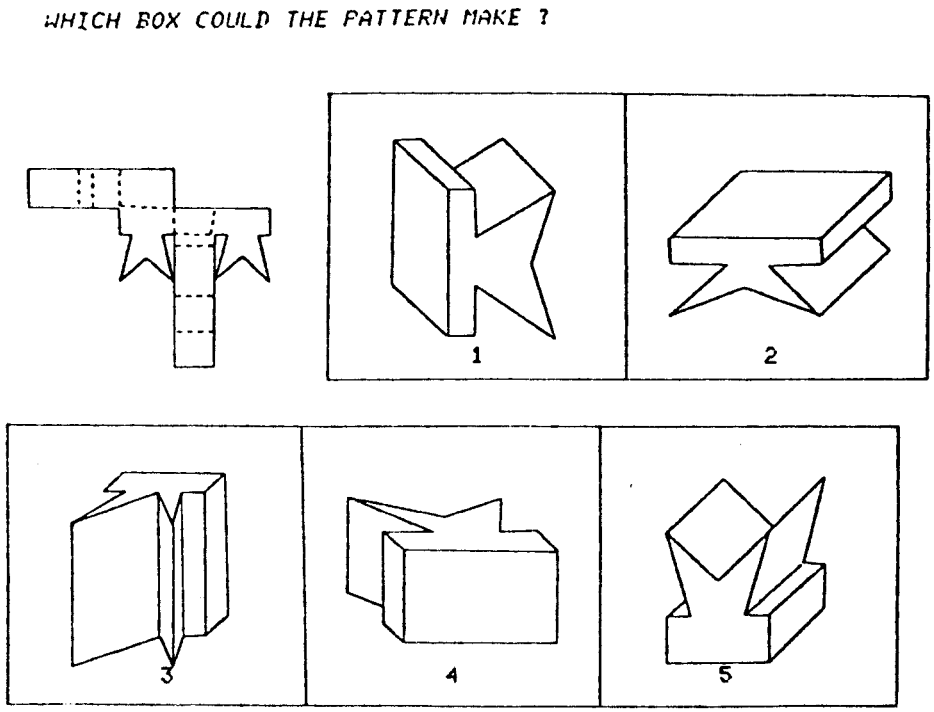
In order to insure proper operation of the system, operating instructions and operating safeguards were built in. The examinee is taught how to use and respond to the terminal before any questions are presented. All solicitations for input are for the characters "1," "2," "3," "4," "5 " and are checked to determine the presence of alphabetic (e.g., ABCD) or special characters (e.g., $!&). If an out-of-range response, an alphabetic character, or a special character is given, the instructions for responding are repeated. Then the screen is cleared and finally the question is repeated. Proper character input is then converted to its equivalent numerical form and processed.

The characters for the questions of WK and AR are kept on an external randomly accessible file and read as needed. Screen control characters are stored with the literal characters, which makes for simplicity of operation. The last array of the file, roughly equivalent to the last record of a FORTRAN file, contains a 4 by $N$ ($N$ = the number of items in the file) matrix of the item parameters and answer keys. This matrix is read in and manipulated prior to the presentation of all questions.

Producing the pictorial displays for the SP items presented a unique problem in storage and drawing. One proposed storage method was to use back screen projection from a random access slide projector. This was discounted because it allowed only about 100 items to be stored, added a mechanical component to maintain, and required photographic slides of each item. Similarly, the idea of writing a specific mathematical function to generate each individual figure was dismissed because it required extensive programming for each new item. Finally, a method of display was developed using the sophisticated graphics capability of APL which only requires placing a drawing of the SP item on a "digitizing tablet" and touching various points on the drawing. These are then transformed into a vector for each item, and the graphics package draws the numerous lines represented by the vector at a very rapid rate (see Figure 1). At 1200 baud, the figures almost flash onto the screen as the vector is read in from its place on the file. This technique could be extended to any item requiring drawings, such as mechanical principles or block counting.

All technology and software developed to draw the Space Perception items are general enough to be used in other testing or educational applications. The perspective and three-dimensional effect are very good, and motion for rotating or shifting the figures can be added. Rotating figures, demonstration

Figure 1
Direct Copy of a Typical Space Perception Item From Screen of CRT




of mechanical principles or moving lever arms may lead to new item types not amenable to static paper-and-pencil tests. Computer driven graphics may enable measurement of new and important ability areas.

An operator's manual, algorithmicly written, was produced for the AFEES personnel. It contains complete instructions for initial daily starting and stopping of the testing system; it also gives instructions for starting the program if the terminal is already running. The manual offers names and telephone numbers of people to contact in the event of trouble. The programs have been "locked" to the AFEES staff, and they have been advised not to try to edit the programs. Back-up copies of both the programs and the files are stored on-line and require only a command from the proper user to reinstate damaged programs or to update programs as they are refined.

Data grade telephone lines, a special telephone number for the AFEES use only, and a special sign-on code were provided to reduce competition for telephone ports in the time-shared environment. The "Special Testing Room" at the AFEES used to house the terminal is a 10' x 12' windowless room containing several student chairs with arms, one side chair, and a 3' x 2' table for the terminal. The terminal and the telephone connector need little space and can be operated in any room with 117 volts AC current and a telephone.

## Demonstration and Future Implementations

The feasibility of adaptive testing will be investigated in the San Antonio, Texas, AFEES demonstration by the assessment of two important factors. First, did the system run with little trouble and attention? This will be assessed from interviews with the AFEES staff and from daily logs of the system's operation. Secondly, was adaptive testing as valid as paper-and-pencil testing? The validity of the adaptive testing system will be assessed by comparing the subjects' adaptive scores and the subjects' AFQT subtest scores. Analysis of these data will help in making future decisions about adaptive testing.

Before any large scale implementation can be undertaken, there will be questions to answer subsequent to the demonstration. Some of these questions are psychometric, some logistic, and some economic. As yet, no testing configuration, either local or nationwide, has been developed, nor have system costs for implementing, operating, and supporting adaptive testing been established. Basic conceptual questions dealing with such diverse topics as testing models, back-up systems, operating policies, and central versus dispersed processing remain unanswered.

It is conceivable that certain other decisions will facilitate broad scale implementation of adaptive testing. For example, the AFEES in Baltimore, Maryland, already has computer-automated management and paper handling on an in-house mini-computer. The addition of adaptive testing might require little additional hardware; and, in quantity, this additional hardware might be inexpensive enought to merit its use. Furthermore, adaptive testing could add to test security because neither test booklets nor answer key are distributed, and no one can have knowledge beforehand regarding which questions will be administered to a subject.

In the future, the actual costs and benefits of adaptive testing will be known. This will permit realistic decision-making for its use. This knowledge will allow adaptive testing to move from a research topic of the 1970s to an operational tool of the 1980s and beyond.

## References

Gilman, L., & Rose, A. APL 360 an interactive approach. New York: John Wiley & Sons, Inc., 1970.

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

Owen, R. A Bayesian approach to tailored testing (Research Bulletin RB 69-92). Princeton, NJ: Educational Testing Service, 1969.

## Acknowledgements

The views expressed herein are those of the author and do not necessarily reflect the views of the United States Air Force or the Department of Defense.