Computerized Adaptive Testing and Personnel Accessioning System Design

Mark A. Underwood
Navy Personnel Research and Development Center

Psychometrics has become increasingly dependent upon computer technology. This is evident in the widespread use of packaged statistical analysis programs, especially the Statistical Package for the Social Sciences (SPSS). Psychologists, in particular, are reared in the environment of the large-scale computer (e.g., IBM 370, CDC Cyber series, Honeywell 6000, Burroughs 6700, PDP-10, UNIVAC 1100/xx), which is amenable to manipulation of large data areas and large programs with appetites for both input/output (I/O) and central processor resources. It would be safe to say that the computer that could adequately satisfy the requirements of the largest statistical programs has yet to be designed. Many of the habits psychometricians have acquired in large computer system usage, however, cannot be tolerated on smaller systems; yet, it is the smaller computer system that will likely provide the apparatus for adaptive testing where it is needed. Therefore, this paper discusses the implementation of adaptive testing and related applications upon small computer systems on a widespread scale.

Background

The Navy Personnel Research and Development Center (NPRDC) has been studying the implications of distributed computing technology for personnel accessioning system design. The availability of lower-cost computer systems makes it possible to configure multi-computer networks which both increase local computing capabilities and reduce network operating cost, especially in the area of telecommunications. Under the mantle of Project CONTRACT (Computerized Navy Techniques for Recruiting, Assignment,Counseling, and Testing), a complex network design has been intensively investigated. This design consists of both fixed and mobile computer systems. These two computer system types are distinguished not only by the requirement of portability, but also by their architectures and capabilities. The mobile system is the one most relevant to the present discussion; it has been configured to perform adaptive testing functions. It is designed also to provide computerized career information, such as NPRDC demonstrated at a high school in San Diego. It is assumed that testing may eventually occur not only at the Armed Forces Entrance Examining Stations (AFEES) but also in the field--at high schools, community colleges, and,on special occasions,shopping centers and exhibitions.

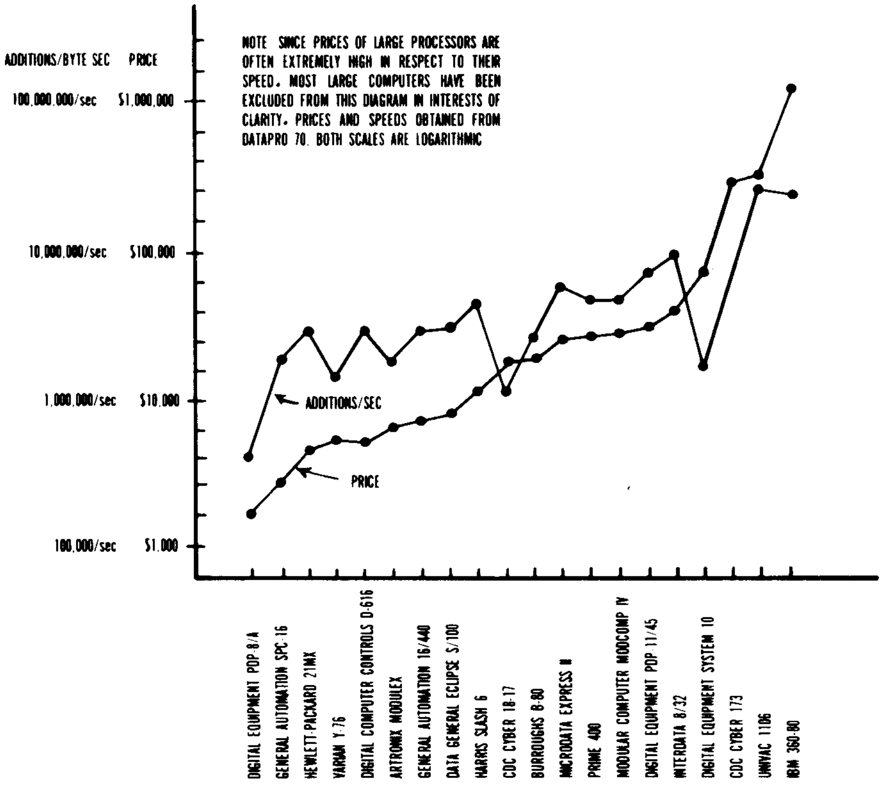
Field testing of this kind requires mobility, because the cost-effectiveness of a computer system physically situated at any one location would be lower than one which could be moved to the location of the next group of people to be tested. Further, the mobile system must be capable of more than adaptive testing. Information dispensing (as well as collecting, in which category testing falls), counseling, quotas, shipping schedules, policies, and perhaps even some criterion-referenced measures are all processes that must be accomodated by the computer system. The mobile system can be interfaced nightly with fixed station computer systems so that local recruiters can follow up on information gathered in the field. Of course, the mobile system need not be mobile; it could be placed in a small office space, such as would be the case at an AFEES.

There are two obvious dimensions for assessment of candidate system configurations: cost and performance. In the course of this discussion, other dimensions will be introduced. However, these two obvious dimensions are important configuration parameters. Cost and performance are intimately related, even with the advent of widespread low-cost microcomputers and their accompanying microperipherals. The breakthroughs have been made in lower memory prices and in the availability of low-cost, low-performance devices which can enable local data collection formatting, editing, and communications. The minicomputer was at the forefront of this trend in the first part of this decade, and the microcomputer will probably obtain the most visibility in the second half. Matching adaptive testing and career information applications to appropriate hardware configurations inevitably calls for compromise between cost-conscious systems analysts and psychometricians; an exceedingly close match between configuration characteristics and applications program resource requirements is essential.

Figure 1 shows the speed vs. entry level system cost of increasingly fast computer systems. There can be no doubt about the possibility of overkill; a program which does not require instantaneous "turnaround" makes use of the same hardware as those which do require it. Programs which are mostly I/O oriented may run no faster on machines with extremely fast central processing units (CPU) because CPU time is not a primary determinant of time to complete execution. The possibility of underkill is obvious; the job is not finished before it is needed. The requirement for multiprogramming, made essential by the multiple-user nature of adaptive testing, introduces the critical factor of the sophistication of the computer system's operating system software. System throughput and resource utilization intelligence decreases dramatically as does both hardware speed and cost.

The level of software sophistication is quite high. For instance, the area of data base structures required to support multiple ability banks, with the possibility of differential starting points based upon ability measured in prior ability dimensions, could become relatively sophisticated. In the worst case, the data base would have to be a fully inverted file, with the key fields the actual raw item parameters as well as item numbers for keys. Furthermore, multiple banks (or files) would have to be interrelatable upon demand.

Figure 1
Selected Entry-Level System Speeds vs. Costs



NOTE SINCE PRICES OF LARGE PROCESSORS ARE OFTEN EXTREMELY HIGH IN RESPECT TO THEIR SPEED. MOST LARGE COMPUTERS HAVE BEEN EXCLUDED FROM THIS DIAGRAM IN INTERESTS OF CLARITY. PRICES AND SPEEDS OBTAINED FROM DATAPRO 70. BOTH SCALES ARE LOGARITHMIC

On the Use of a Minicomputer

Hardware and software performance monitors are techniques for studying some parameters of program behavior. If a well-written program is analyzed using such tools, quantitative information about computer resource use can be gained. For adaptive test administration software, statistical distributions of machine instruction type can recommend certain architectures. For instance, the occurrence of floating point hardware instructions would indicate the possible value of specialized hardware for floating point--beyond the "floating point boxes" provided by the minicomputer vendors. In a more sophisticated analysis, such studies can reveal patterns of I/O, memory referencing, and class of machine instruction. Hardware and software monitors are an integral part of the effort of matching system functional specifications to hardware and software characteristics.

System requirements. The rationale for the selection of a multi-minicomputer design for a mobile counseling and testing system is worthy of discussion because of the variety of design alternatives available. In addition to the low-cost motivation, these factors led to the identification of a system configuration with the following characteristics:

1. Need for computer control of the graphic items;
2. Capability for performing the adaptive sequencing algorithm locally without deferred scoring or recording;

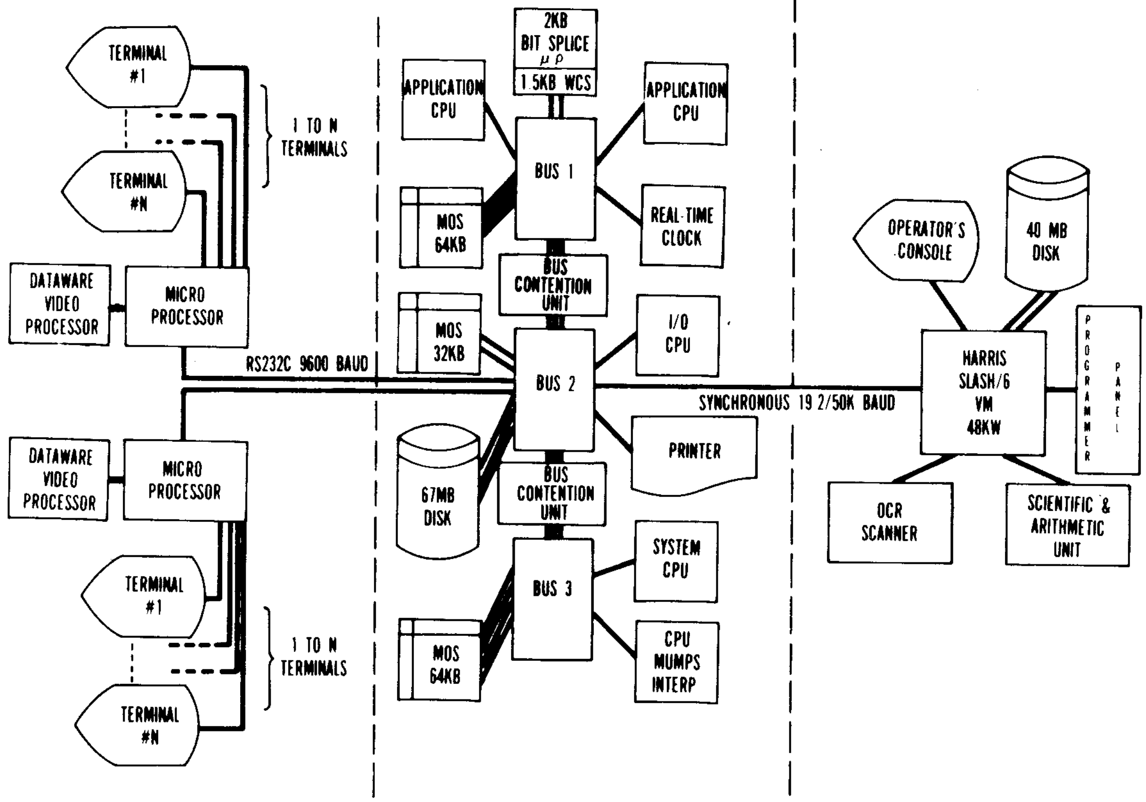
3. Capability for tracking sequences of polychotomously scored items;
4. Increased requirement for test security, allowing for a large item pool to reduce the probability of two individuals with equivalent ability and/or educational backgrounds receiving identical items;
5. Reduce distractors by presenting only one item at a time in the stimulus environment, when necessary;
6. Capability for a large multiple-bank item pool with a complex data structure for rapid retrieval and sophisticated ordering algorithms;
7. Capability for updating the ability estimate after every item presentation, using the most sophisticated strategy (a worst case situation, in terms of computing requirements); and
8. Considerable expansion capability in terms of hardware and software features and also in terms of the number of sophisticated algorithms for stimulus presentation and response data collection and analysis.

Alternative hardware. Utilization of off-the-shelf hardware and software, where performance and cost were in the orders of magnitude needed, simply has not arrived in a sufficiently low-cost package to warrant incorporation into adaptive testing stations. Whether this will change depends upon developments in artificial intelligence and an improved understanding of semantic processes. Even then, it can be certain that this will result in the implementation of highly sophisticated algorithms for language processing which would add significantly to an already burdened small computer system that was chosen for its relative simplicity of design and low cost. The area of video image processing seems to hold the most promise because it presents a means of automatic image digitization and manipulation without excessive expense.

In addition to the high probability of mechanical failure and limited storage capabilities in some cases, microfilm and slide image presentation packages lack the dimension of computer-controlled image manipulation and storage; sensory psychologists would surely certify the importance of control of the image. A look at current item banks shows that photographic-quality images are not frequently used; for instance, a drawing of a vise is used instead of a photograph which would more closely resemble what would be encountered in a military shop. The technology described possesses the capability of presenting "snapshots" from video cameras; a text or discussion could accompany snapshots, or items could consist of sequences of snapshots. The implications for item construction are significant and many. Thus, both hardware and software must be assessed in the identification of a suitable architecture for adaptive test and career information computing.

A prototype. Figure 2 is a prototype configuration under consideration at NPRDC which utilizes multi-vendor, multi-processor technology. It is expected to be able to perform many simultaneous test administrations and, at the same time, is capable of some heavy duty statistical and scientific computing. Also, its use in large quantities would place it in the middle to high end of the mini-computer category. Note that specific functions have been allocated to the various processors in the configuration, as indicated by labels in the processor boxes. Unlike the PDP-11 UNIBUS structure, the Modulex system is capable of mulitiple bus structures as well as multiple processors, and it has twice the rated bus speed. Therefore, it is less likely to become bogged down from movement of data on the bus between processors, from memory to processor, and from peripherals to memory or disk. Aside from the not insignificant task of multi-bus and multi-processor coordination, however, the Artronix operating system is unremarkable, particularly with respect to memory management and higher-level languages.

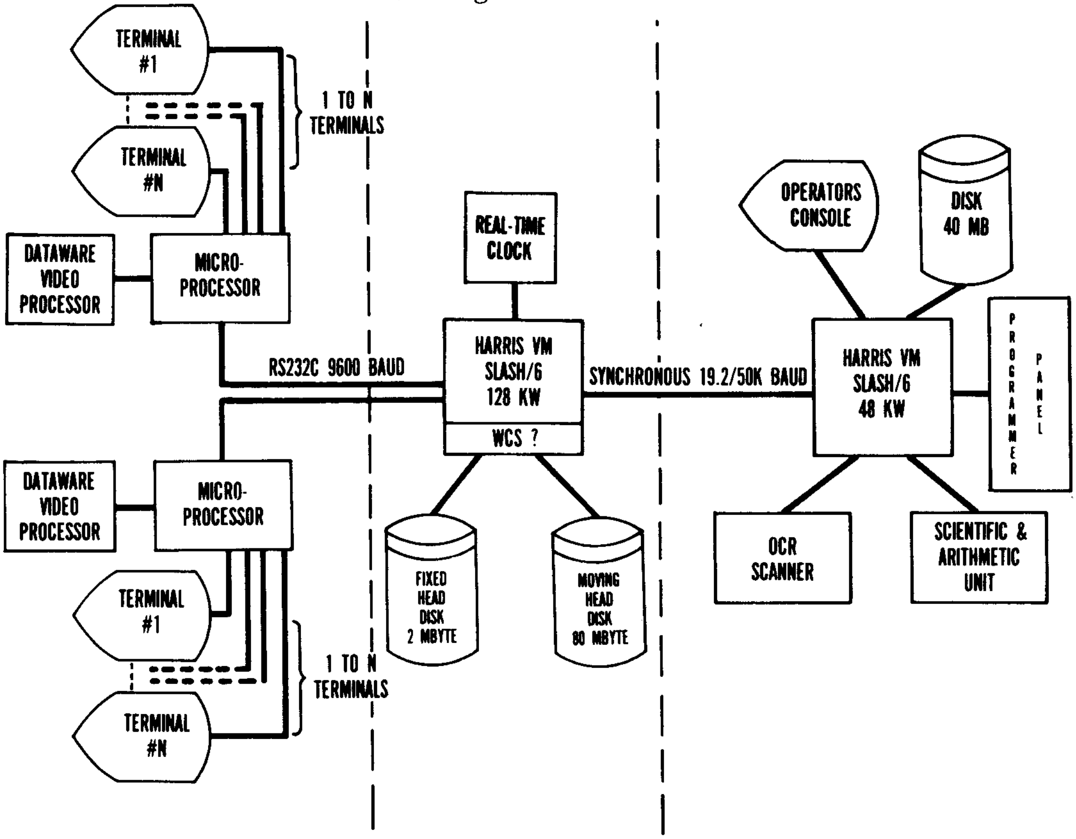
Figure 2
Configuration A



The prototype can be broken down into three functional parts which have been named for convenience the "front-end," "middle," and "back-end" subsystems. The Artronix Modulex, or a similar system, will comprise the middle. The actual applications code which supports the administration of test items will, in large part, be contained in the middle system, and most file maintenance will be performed there. The back-end system will closely resemble the Harris/6, a virtual memory computer with a 24-bit word size and a parallel scientific arithmetic processor unit. The Harris/6 will be responsible for the number crunching, notably floating point number crunching, and array processing associated with adaptive testing, assignment algorithms (if implemented in the mobile environment for recruiters) and career information analysis. In .the front end will be located several microcomputers which are responsible for handling the graphics display information, decoding, transmission of limited-keyboard information, polling of many devices, and selection of appropriate station or information received by it from the other subsystems.

Other configurations under consideration are shown in Figures 3 and 4. These configurations represent differing emphases upon the I/O or computing requirements of a live testing environment. Obviously, a final selection cannot be made without considerable further study of such areas as resource utilization, component reliability, and cost trends.

Figure 3
Configuration B

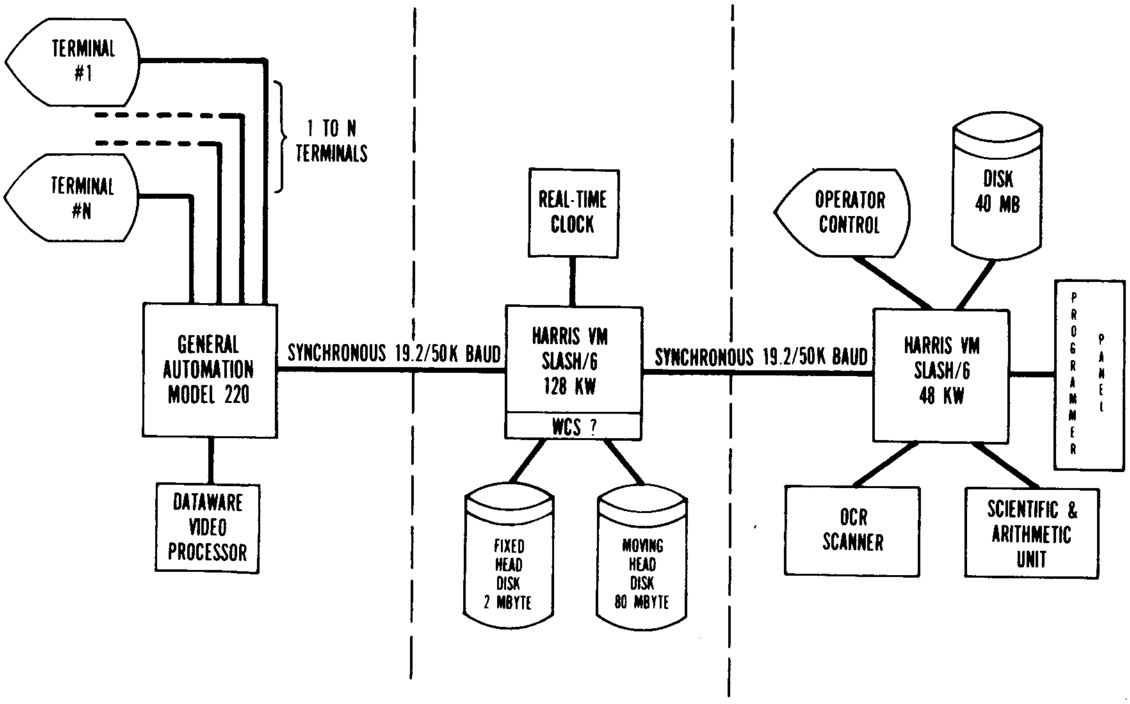


Considerations in Testing System Design

Item Presentation

Those who have considered adaptive administration of military tests may consider the Armed Services Vocational Aptitude Battery (ASVAB) as one possible item pool source. With this in mind, the items in the ASVAB were examined to determine the computer resource requirements of item presentation. Resource utilization occurs with respect to item presentation along several dimensions: (1) storage, (2) display formats, (3) system response time, (4) number of examinees to be supported, and (5) examinee response measurements. Economies along these dimensions can directly affect test construction and evaluation strategies, as well as the cost and performance of the supporting computer configuration.

Figure 4
Configuration C



"Front-end" clusters. The best estimates and evidence of the Data/Ware 12/40 image-processing system show that a high-speed microprocessor is capable of expanding compressed video images and displaying them at as fast a rate as 7.5 256 X 256 dots or "pixels" per second. This clearly demonstrates the feasibility of providing high-performance graphical item presentation for a large number of terminal-type devices with a single microprocessor. As the performance information about adaptive test image presentation becomes available from prototype systems, estimates can be made concerning the number of devices that can be supported by a given hardware configuration. The number which can be supported is a function of the available memory for data areas in the "middle" subsystem, the amount of floating point arithmetic and array processing which must be performed, the percentage of graphical items presented, the rate of progression through the test by respondents, and the speed of communications with the middle and back-end subsystems. Additional overhead must be allowed for performing poll-and-select operations upon data coming to and from the terminal-type devices or the test stations.

Storage. Moving head disk drive and controller assemblies on mini-computers cost between $8K and $40K, depending upon total storage capacity and rated transfer speed. At a cost of about .045¢/byte, or .0057¢/bit, low-speed storage on high density drives may seem relatively inexpensive when compared to floppy disk drives, for instance, at .64¢/byte. However, the expense of such drives on mobile systems represents as much as the cost of the CPU and a considerable amount of main memory. The requirements for storage must be carefully estimated and monitored while the system is in operation. Storage of item information consists of item parameter data and the item itself; the storage needs of the latter are of particular interest.

Existing computer terminal technology provides for increasingly lower-cost display of standard 5X7, 7X9, or better dot-matrix alphanumeric symbols. While the low end of the cost per unit with industry standard bare bones features seems to be around $850/unit, this is not likely to decrease, although the number of available features may increase due to increased use of microprocessor-based terminal designs.

This technology would handle roughly 60-75% of the existing ASVAB item pool and with the use of intelligent compression strategies for textual data could occupy as little as 480 bytes times the number of items (200 items = 96,000 bytes), or about $43 of disk storage excluding system overhead. However, increased item pool sizes would be desirable for such reasons as security, increased test resolution, and sophistication. It seems to be not unreasonable to assume the need for a very large item pool for the purpose of uncompromisable, comprehensive ability assessment for military personnel accessioning. The storage would range from 1.2MB to 2.4MB and cost around $1,080 for the large drives. Utilization of available space would be at about 3.5%. This would rule out the use of floppy disk drives and smaller cartridge drives.

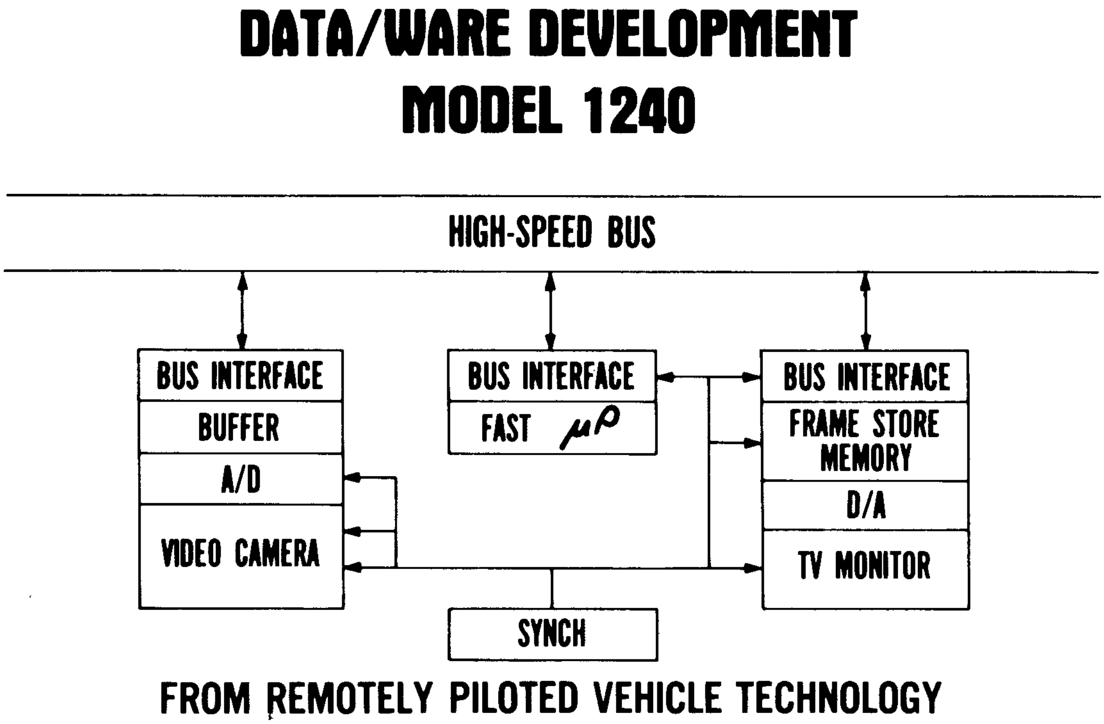
The test administration programs themselves would not require more than 1.5MB storage, including dynamic data areas. One adaptive testing program studied utilized only 16 bytes per item for psychometric information (item difficulty, discrimination, guessing coefficient); and this information would consume only another 40-80KB. However, it may be desirable to store additional item information at this level for on-line use of graded response models (Samejima, 1975).

Display formats. Technology is available for presenting conventional alphanumeric information on a CRT, and the cost per unit could be reduced by minimizing capabilities. However, the presentation of pictorial items is another matter. Plasma, microfilm, and conventional graphics terminals presently cost between $5,000 and $80,000 depending upon capabilities. At $5,000 per station, adaptive testing could never become cost-effective. Therefore, an investigation of military research and development into image processing techniques was made. The strategy used in remotely piloted vehicles seems to hold the most promise. This technique involves (1) the use of high-speed microprocessors and image compression methods to convert analog conventional video signals to a digital format for encoding for either storage or analysis by computer; (2) transmission by telemetry in digital form; and then (3) ground reconstruction back to a video format through decoding of the digital signal and the use of a video gray scale.

The technique will be studied for applicability to adaptive testing by a pending contract with NPRDC. The appeal of this approach is manifold: (1) it can reduce the cost of graphics-capable stations to well under $1000/station; (2) it provides for a built-in image compression scheme which will conserve storage of items in the system; and (3) it gives psychologists better control of the stimulus configuration by providing it in a computer-manipulatable format. Naturally, it will be possible to superimpose alpha-numeric and graphical information upon a single image. Figure 5 outlines the way a currently operational video-based graphics system works as developed by Data/Ware Development under contract to NOSC and Wright-Patterson AFB.

Even utilizing dynamic compression techniques, storage of graphics items for adaptive test administration requires considerable storage on disk. It has been estimated that a typical item would require as much as 50K bits, depending upon the nature of the image(s). This would require at least another 1.5MB on disk. Further, because of the availability of this capability, the number of graphic items may increase for adaptive testing applications. Not nearly enough is known about the parameters of such processing, and the pending contract will address these issues in a more quantitative fashion, as well as setting out hardware and software specifications for a prototype system with these capabilities.

Figure 5
System Block Diagram of Data/Ware Development
Video Graphics System



System Response Time

The penalty for waiting for the computer to respond to a command varies, depending upon such factors as the status of the person waiting and the degree of urgency for the particular task. In the adaptive testing situation, waiting for the system has serious consequences. The most obvious effect is that it increases test administration time, the minimization of which is supposed to be one of the advantages of adaptive testing. Less obvious are such factors as the influence upon examinee attention, the significance of environmental distractors, and anxiety and motivational factors. System response time is also an important dimension in a small computer system; it is even more important because of the relative ease with which applications can bog down the system. In general, systems programming and applications programming must be carefully inspected; and all jobs must be of a carefully predetermined and carefully engineered variety. It is desirable for the job mix, as much as possible, to remain relatively homogenous with respect to resource utilization and program behavior.

Many factors can cause decreases in response time. Probably the most important to keep in mind is the small number of resources. For instance, there is only one disk for general system use; and contention for it will be high, especially if the amount of main memory is kept small for economy purposes.

Number of Examinees to be Supported

In determining system cost-effectiveness, the number of simultaneous test administrations which can be performed is a very important factor. In general, with minicomputers as the number of simultaneous users increases, the cost of the overall configuration rises in direct proportion. Analysis has shown that conventional channel multiplexing techniques for data communications (usually RS232C ports) would be an inefficient way of communicating with a large number of test stations. A design objective has been to develop a capability for as many as 64 simultaneous administrations. Whether or not this is feasible with the contemplated design is difficult to answer at this time. Table 1 shows some of the parameters influencing the number of stations which can be serviced in a reasonable response time. Sixteen administrations is a lower bound for system cost-effectiveness, i.e., simultaneous administration of fewer examinees would reduce system cost-effectiveness to an unacceptable level. While the AFEES environment may not presently require so many, a high-school administration would have to occur on a very large scale in a relatively short period of time in order to compete with the economy of scale offered by conventional paper-and-pencil tests.

System costs can be broken into fixed and variable cost factors. In order to administer only 16 tests, a certain minimum system is required. As the number of examinees increases, memory and processor use are utilized in direct proportion to the number of administrations and the amount of disk concentration goes up by a more complicated formula. Cost-effectiveness assessment for this kind of configuration is difficult to perform; and it is necessary to know data external to the computer system performance, such

Table 1
Factors Influencing Number of Serviceable Stations

Speed of Dataware Video Processor
I/O Capacity of Microprocessor
Main Memory Size of Microprocessor
Speed of I/O Interface from Front End
I/O Buffer Size of Front End
I/O Handling Speed of I/O CPU
Space and Electrical
Back-end Computing Speed
Size of Code and Data Areas for all Programs

as average test length under adaptive testing in order to arrive at better estimates. The value of field tests in this regard cannot be exaggerated. The desirability of large-scale administrations also depends upon policy factors and the working relationship with public institutions for which testing may be offered as a diagnostic and/or vocational service. Although direct comparison with the present ASVAB program is tempting, the quantum step forward represented not only by the computer system itself, but also by the adaptive testing strategy, suggests that comparison would be inappropriate.

Examinee Response Measurements

The adaptive test program studied utilizes a dichotomous (correct/incorrect) response structure. There are other response measurements which can be made. For instance, time-to-respond, or latency, could be collected. For graphic items, relative image size and resolution may be relevant; an individual "calibration" might be required, based on the conclusions of basic research into the relationship between stimulus characteristics and examinee performance on an item.

Indices of Testing Cost-Effectiveness

Issues of cost-effectiveness will be raised more than once before adaptive testing is widespread in its application. Contributing to the confusion about the cost of adaptive testing will be the fact that hardware (i.e., primarily computer hardware) is involved in adaptive testing, whereas conventional paper-and-pencil testing is essentially labor-intensive. Conventional testing is costly in terms of examinee and proctor hours and scheduling of space and people; however, these items are not directly chargeable and represent consumables rather than capital investments. In a technologically advanced culture such as ours, it is relatively more straightforward to develop algorithms for justifying hardware investments where labor savings can be shown. The ease with which automated testing techniques can be "sold" will depend upon the consequences of the current inconveniences, inaccuracies, or difficulties associated with paper-and-pencil testing. In military personnel acquisitions, the problem is also one of market penetration: to test as many people as possible in as short a period of time to minimize disruption to the host institution's environment. Some of the parameters which influence adaptive testing cost-effectiveness are:

1. Mean response time of examinees to (video-presented) items;
2. Computer system configuration size;
3. Mobile support subsystems cost (if applicable);
4. Maintenance costs for hardware and software;
5. Training associated with staffing of computerized adaptive testing stations or vans;
6. Number of items required to perform necessary ability measurements;
7. Capability of computer system to perform other recruitment-related tasks (e.g., computerized career information systems, data management, individual job assignment), with associated cost-reducing effects;
8. Cost and speed of data communications (very little is involved in the configurations proposed in Project CONTRACT);
9. Number of hardware items procured within an 18-month time frame;
10. Estimated system life for hardware and software;
11. Installation costs for testing stations;
12. Research and development: computer science, psychology, operations research;
13. Hardware and software quality and reliability;
14. Number of simultaneous test administrations; and
15. Frequency and cost of item calibrations.

A sample systems cost is presented in Table 2. Relevant indices are cost per item administration, cost per test, number of simultaneous administrations permitted, and daily operating expense. Obviously, the number of items required to make ability estimates and the complexity of testing algorithms employed are primary factors in the determination of cost effectiveness.

Problem Areas

Systems analysis of the configuration alternatives for adaptive testing and of the software techniques presently in use for adaptive testing have suggested some problem areas that are worthy of attention. This attention must be given jointly by computer scientists and psychometricians if a cost-effective, fully functioning system can be delivered in a reasonable time-frame. At best,topic areas can be delineated in the present discussion; further research is obviously needed.

Frequency of estimate update. Estimated mean and variance of a normal Bayes prior distribution, given observation of dichotomous response to an item with known characteristics, can be updated on an item-by-item basis or after a set of items have been answered. CPU (though not I/O) demands could be minimized through the use of item sets where possible. Item set size could be a function of the amount of precision required in the ability assessment, logical or substantive interrelation, item structural similarity, or simply a convenient "page" size for use on a video screen.

Table 2
Sample Systems Cost (Estimated)

Item	Cost
Computer System @ 50 unit quantities	
Includes: 5 independent memory "banks"	
Six minicomputers	
110 MB moving head disk storage	
Matrix printer (500 cps)	
OCR	
Front end microprocessors	210K
Video Devices (64)	32K
Van Excluding Power and Air Conditioning	30K
Power and Air Conditioning	10K
Total	282K
*Maintenance (.5% of purchase per month, serviced by government personnel)	700/mo.
*Transportation (10K miles per year, vehicle maintenance, etc.)	300/mo.
*Communications	45/mo.
*Training and Miscellaneous Software Maintenance	100/mo.
5-year Total Recurring Costs	68.7K/5yr.
Installation	2K

Assumptions: 64 stations, 70% overall utilization;
School calendar (180 days) in use;
35 items to criterion for adaptive testing;
Item presented average of every 20 seconds;

Over 5 years, could administer: 1,240,000 test @ 28¢ per test**

* Probably would be offset by manpower savings through improved procedures in recruiting.

**"Test" assumed to be one sequence of adaptively administered items whose mean length is 35 items.

Pool size. An important topic is the implication of item pool size upon test reliability, precision, security, and administrative efficacy. Pool size affects the storage requirements and, if it reflects a more sophisticated item-sequencing algorithm, increased CPU usage. Items can have different "versions" or possess redundancy along substantive dimensions. A data base structure which reflects such relationships must be designed to allow for intelligent access to item raw data by applications programs.

Precision. One problem with many small computer systems is that they are not designed for high-precision applications. This is frequently a function of the common 16-bit word size; even double precision floating point arithmetic can result in insufficient computational accuracy, perpetuation of roundoff errors, truncation, and loss of significant digits. This problem can be exacerbated if the computing for sequencing moves at all in the direction of a convergence failure. A parallel-processing floating point array processor may be indicated if conventional mini-architectures are not adequate to keep up with the demands of this applications area. Also, given that parity errors do occur in disk and main memory systems and because losses of single bits can have disastrous results in the computation of statistical algorithms, this must be provided for in advance. Hardware must be specified which has main memory byte parity and correcting and checking.

Storage of sequence information. As each examinee traverses through an adaptively administered test, a probable unique sequence pattern is generated. This pattern is a form of response data which may or may not be utilized by the adaptive algorithm for administering the next item(s). As this sequence grows lengthy, the overhead associated with maintaining that information and incorporating it into analysis becomes significant. The tradeoff between the precision gained by such sequencing data, if any, and the associated computing overhead should be estimated; such sequencing data may have some value in evaluation of the algorithm and the testing process as a whole.

Possible use of differential starting values. It has been suggested that externally determined criteria for the starting point in an adaptive test administration may be utilized. This significant possibility raises the idea of an integration of the career information aspect and the testing aspect of the mobile computing system for recruitment. Counseling-type information may suggest starting points. Other test scores may suggest starting points, or recruiters might make use of the adaptive test capability to perform screening (in which administrations would not be carried to completion). The interrelationship of programs and data bases for these various application areas must be carefully studied to determine if dynamic task synchronization should be made another requirement of the target system's operating system software.

Use of FORTRAN. The acceptance of FORTRAN for the implementation of adaptive testing algorithms has both advantages and disadvantages. A prominent advantage has been that psychometricians have been able to write some or all of the code. For the implementation of statistical algorithms, excluding data manipulation, FORTRAN is adequate and efficient on small computer systems. However, its use for any I/O--especially disk I/O--is contraindicated. This can be said for most of the FORTRAN implementations on small computer systems. FORTRAN is especially poor for formatted Input/Output because it uses excessive CPU resources and makes poor use of memory while doing it. Access to complex file structures, too, may be cumbersome. The use of FORTRAN must be restricted to those domains in which it generates efficient and controllable code for small computer systems. Certainly, only part of the adaptive test administration process can utilize FORTRAN, for its reckless use can result in needless proliferation of CPUs and memory.

Requirements of Adaptive Testing Systems

Program Attributes

An important means of determining appropriate hardware and software requirements for adaptive testing is to study the characteristics of existing programs. Examination of these FORTRAN programs reveals useful information which is presented in Table 3. Separation of code and data areas by some minicomputer compilers (an important capability) permits direct inspection of the code and data requirements for each routine, including shared COMMON data areas. Program code, or instruction space, is the only program memory resource which can be shared between multiple executing users (or, in this case, examinees). Therefore, the ratio of code to data memory resources needed by these programs is an important characteristic. In the item administration program summarized in Table 3, a maximum of 24% of the memory resources needed can be shared between users. In a minicomputer environment with a target support capability of 64 simultaneous users, this is unacceptable; a total of 749,312 bytes would be needed for data areas in the worst instances.

Table 3
Program Attributes of a Bayesian Ability
Estimation Program (Simulation of Actual Sessions)

User Code Segments: One Main, Two Subroutines

1. 2364 bytes
2. 165
3. 252
4. (FTN Intrinsics) 902

TOTAL: 3683

User Data Segments

1. 10755
2. 72
3. 80

TOTAL: 10907

Ratio of Code/Data Areas for Program as a Whole:

3683/10907 = .338
Data Areas Represent 74.8%

File Space (if all files open): 801 bytes
Minimum Memory Required by B1700: 3643
Dynamic Area Required for all Data Pages to be in Memory: 14216
Percentage of Sharable Memory Space = 23.9%
Execution Time on B1700 for 32 Simulated Subjects with Maximum 64KB Main Memory, and 16 Items = 5 min. 12 sec.

On the NPRDC minicomputer, a 64KB Burroughs B1700, the operating system's virtual memory capabilities make it possible to execute programs for which the data and/or code space requirements are greater than what is actually available on the computer system. This is accomplished by breaking up the code and data areas into variable size "pages," which can be written to disk and then brought into memory when needed and/or when space becomes available. The Burroughs B1700 FORTRAN compiler identifies at BIND time its minimum dynamic memory requirements. For the item administration program, only 3643 bytes of the total required 15391 data, code, and buffer areas need be present in memory for the program to execute. However, the penalty to pay for this capability is the overhead associated with monitoring page references, reading and writing pages to disk, and producing compiler output that is pageable. The generally slow speeds associated with these techniques are not adequate to the task of high-performance test administration/adaptive testing while simultaneously performing counseling and other application programs. The only solution is to improve the way in which the programs are written and to configure hardware that is exceptionally well-suited to the job at hand. Table 4 shows some common programming flaws in the software studied. As can be seen, a number of poor programming techniques which can severely reduce testing throughput time are evident.

Table 4
Some Poor Programming Techniques Found in Item Administration Programs

Every iteration of some routines recompute items that could be set in DATA or computed just once at execution time, for example :

```
    PI = 3.14, etc.
    C1 = 1.0/PI
    C2 = 2.0/SQRT(PI)
    SIGN = -1.0
```

No use of tabbing on formatted output

Excessive use of FORTRAN formatted I/O (very CPU-consumptive)

Excessive use of GO TO's; unstructured logic is common; makes it difficult to optimize code

Inefficient array referencing and dimensioning

Successive IF statements not placed according to frequency of branch (where applicable)

Space-wasteful ordering of variables in BLANK COMMON

Other Requirements

Use of firmware. The advent of user-microprogrammable minicomputers (e.g., those by General Automation, Hewlett-Packard, Varian, and Computer Automation) has made it possible for particular repetitively executed primitives, which are very low-level software constructs, to be coded in firmware. Some of these firmware capabilities are evidenced in the Hewlett-Packard 21MX mini on which a number of FORTRAN-generated primitives are available in microcode. Possible areas for microcoding for the adaptive testing algorithm are: absolute value function, min and max, float, fix, Pearson's *r*, modulus, passage of subroutine arguments, and multiple dimension array references.

Hard copy output. It is not clear at this point what the requirement for hard copy output at the end of an adaptively administered examination might be. As the need for high-speed hard copy could represent a considerable investment, some thought should be given to whether this is a requirement, and if so, how fast hard copy must be produced and in what form.

Test calibration and propagation. This treatment of computer requirements for adaptive testing has not addressed the issue of test calibration, which is currently performed primarily on large scale computer systems. A nationally utilized test could need field updates or new calibration data. The frequency of such update should be estimated. It should also be determined what data should be collected by field systems to aid in refinement of instrument precision. The use of the computer for test administration should aid in continual improvement in the test merely through simplifying data analysis procedures and facilitating field implementations of new versions of the test pool or algorithms.

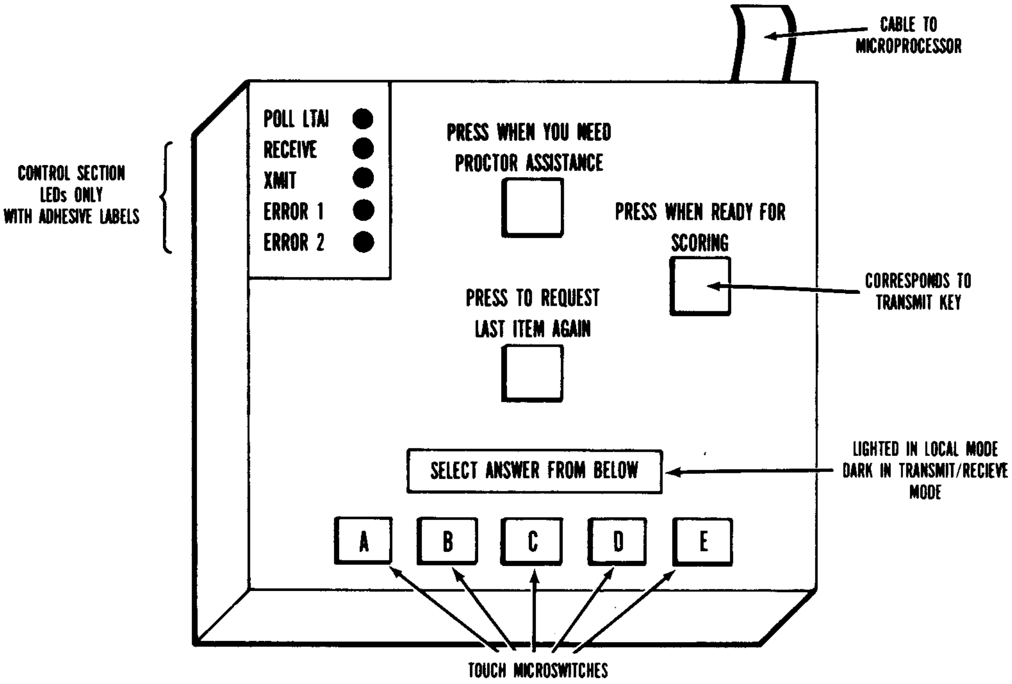
Recovery, reliability,and cross-checking. The possibility of catastrophic failure always exists in non-redundant computer systems. Total hardware redundancy for the system under consideration was not judged to be a cost-effective additional capability. Therefore, recovery from software (either caused by systems software or applications software failures) is a contingency which must be planned for. Complete automated test data generators should be used (such software is already available on larger computer systems and could be adapted for use on small systems) to exercise the software fully before it is pressed into full use. Whenever possible, software should have built-in cross-checks to assess its proper functioning. A valuable heuristic approach is to assume that physical system security has been compromised. File protection (i.e., protection of data and programs) is as important as physical system protection. Simple errors such as channel cross-talk could result in mix-up of item presentations with responses; this is a phenomenon sometimes observed in many timesharing systems when pointers are lost or files are partially destroyed.

Recovery from hardware failures is also significant. The use of periodic "dumps" is worthwhile so that examinees do not have to totally restart examinations. However, there is overhead associated with all of these precautions; and their cost must be weighted against the value of the protection.

Keyboard design. It is doubtful that a full typewriter keyboard is required for adaptive test administration. The technology developed by the Air Force Human Resources Laboratory (Kirby & Gardner, 1976) shows attention to cost and functional requirements along these lines. Psychologists and engineers must plan a minimum keyboard design which is reliable, easy to use, self-evident in function, and totally detachable for plug-in replacement. The way has been shown by the financial institutions, as special teller keyboards have been designed to meet specific transactional requirements. Some inherent logic should be present, such as illumination of error conditions and of line activity to prevent key-in attempts during transmission while buffers are being filled.

Figure 6 shows a possible minimal keyboard for an adaptive test station. It is generally accepted that five-alternative multiple-choice items work better than those with fewer alternatives. For this reason, the suggested keyboard interface layout is shown with a five-response touch microswitch capability.

Figure 6
Possible Minimal Keyboard for
Adaptive Test Station



Computer Technology and the Stimulus-Response Configuration

Some experts have speculated that the future holds a considerable number of alternatives for departure from typical stimulus and response configurations for test administration. The role which the computer and related peripherals can play in contributing to the increased comprehensiveness of psychomotor, ability, perceptual, and behavioral assessment depends to a large extent upon the extent of dramatic cost-reductions in many electronic components and the strides made in software engineering. The rate of these transformations has not been as great as had been thought. The cost of sophisticated graphics, or image manipulation devices, ranges between $15K and $200K per station. This is certainly not appropriate for individual adaptive test stations. So far as natural language processing is concerned, speech understanding systems lag considerably behind speech synthesis. The computer is very proficient at cost-effective presentation of alphanumeric information, so that presentation of spoken instructions or test items would have to be justified on an affective, motivational basis or increased psychometric sensitivity, which is difficult to make cost effective.

The U.S. Civil Service Commission has stated that a nationwide computer-assisted testing system for federal jobs is being developed for the 1980 timeframe (Urry, 1977). The design of a nationwide network should be undertaken with a serious view to the level of processor which is required to support the chosen adaptive testing strategy and the cost of a network, if one is needed. The systems design NPRDC has undertaken to identify a system capable of the implementation of sophisticated adaptive testing algorithms does not require a nationwide network in a real-time sense of a network. The costs of such a network are extremely high, simply for communications costs alone. A centralized system design is less likely to be cost-competitive. Furthermore, the hope for instantaneous and continual application of extremely complex theoretical/statistical procedures must be tempered by the high probability of needing a costly computer configuration to support them.

Conclusions

It has been the intention of this discussion to raise more questions than it has answered. This is the case because of the need for considerable research involving psychometricians and computer scientists in developing a small computer system equipped with the proper combination of hardware and software architectures, which is priced low enough to make adaptive testing cost effective and with enough power to get the job done simultaneously for a sizeable number of examinees. The use of a prototype computer system is essential even if not all of the possible alternatives can be quantitatively evaluated prior to its procurement. With such a prototype, experiments can be designed, simulations can be performed, and hardware and software configurations can be benchmarked. At this point, the questions are numerous enough so that it is difficult to order them according to their impact on system performance. Despite these reservations, the technology of the minicomputer has come a long way in recent years; and much more is possible with the same hardware than was previously. Determining where off-the-shelf hardware and software is not adequate so that customized engineering development can be initiated is one of the first goals, because of the lead time required for development prior to deployment.

The policy and applications contexts into which the adaptive testing system will be placed are also important systems design parameters. It is generally agreed that a multi-purpose mobile system has broader appeal and greater cost-effectiveness justification than one dedicated solely to testing. Where testing is to be part of a larger computing environment, its resource requirements and program behaviors must be assessed in the complex of the overall system control, especially control of memory allocation. If the need for a comprehensive information-dispensing and information-organizing capability is not immediately apparent, recruiters, with their necessarily urgent needs for meaningful and marketable data, will certainly make known the need for a fully integrated system. The parts to be integrated will include: career policy guidelines, information, some goaling data, a mini job-assignment capability, adaptive testing, logging of psychometric data, and recruiter management information. If such a system were to be delivered next year, there can be little doubt that this list would grow simply because of the added capability that a computerized system represents.

References

Kirby, P., & Gardner, E. Microcomputer controlled, interactive testing terminal development (AFHRL-TR-76-66). Lowry Air Force Base, CO: Air Force Human Resources Laboratory, Technical Training Division, 1976.

Samejima, F. Graded response model of the latent trait theory and tailored testing. In C. L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing, Washington, DC, 1975, pp. 5-15.

Urry, V. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Acknowledgements

The opinions expressed are solely those of the author and do not necessarily reflect official Department of the Navy policy.