THE PRESENT AND FUTURE RELATIONSHIP OF TECHNOLOGY AND TESTING

Frank B. Baker

CSE Report No. 235
1984

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

INTRODUCTION

One of the hallmarks of the testing movement has been its long-term involvement with technology. In the early educational testing movement literature, the multiple choice item was referred to as the objective type item because its scoring involved no judgmental processes. Because of this, it was natural that efforts be made to mechanize the test scoring process. Over the years, many schemes were devised to efficiently score multiple choice tests via manual or mechanical devices of one type or another. While these devices worked, they could not really cope with the volume of test results being generated by the national testing programs established during the 1930's.

The first real technological breakthrough in the automation of test scoring was due to the efforts of a high school physics teacher who invented a current summing device that could obtain a test score from pencil marks on an answer sheet. His invention became the IBM 805 test scoring machine of 1935 that was in widespread use until the 1960's. Many of you are probably aware of the idiosyncracies of this machine.

However, as testing programs grew in size, the need for faster, more accurate automation grew. The second major technological advance, occurring in 1955, was the development of the optical mark reader (Baker, 1971). This device could sense marks on the answer sheet optically and used the recently developed computer technology to

score the test. Much of this scanning technology, developed by E.F. Lindquist at the University of Iowa, resulted in his 1955 MRC machine which could process 6,000 answer sheets an hour.

Since that time there has been a steady improvement in the accuracy, flexibility, and versatility of the optical mark reader (OMR) test scoring equipment. Today, there are a number of manufacturers of such equipment (Baker, 1983). Because of the capabilities and availability of such test scoring equipment, a wide variety of organizations (commercial testing companies, state education agencies, universities, and school districts) can conduct large-scale testing programs in a cost-effective manner. As a result, one can consider the automated scoring and reporting of test results to be a rather mature field. However, it has provided the basic foundation for other applications of technology in testing.

STATE OF THE ART

The high capacity test scoring equipment represents one end of a range of test scoring equipment. At the other is the desk top scanner. In about 1967, Richard Schutz of SWRL explored the possibility of having a desk top scanner developed. However, events conspired to thwart this effort. It wasn't until 1974 that a commercially available desk top scanner (the DATUM 5098 OMR) was available. It could handle a 64 x 13 array of marking positions on an 8 1/2 x 11 sheet of paper, and cost $3,000.

This scanner was important because it could be connected in series with a computer terminal and item responses could be sent over

telephone lines to a central computer for scoring and reporting. The desk top scanners opened the door to cost-effective, small scale, automated test scoring and reporting. At the University of Wisconsin, we have DATUM scanner serial number 3, and it is still serving us faithfully for classroom testing. The DATUM 5098 was a very basic OMR and did little more than sense marks on the sheet and send item responses to a computer with an overall processing rate of about 15 sheets per minute.

In recent years, the microprocessor chip has become readily available, and desk top scanners have begun to incorporate them. The Europeans have been a few years ahead of us in this area and small scanners such as Kajaani Evalmatic use a microprocessor to compute and display the test score as the sheet is scanned. Recently, desk top scanners incorporating microprocessors have appeared in this country as well. Because these microprocessors are small but fast computers, they can be used to perform a number of functions within the scanner, one of which is the quality control of the mark sensing process. The microprocessor can obtain the readings from multiple marks, apply decision logic, and ascertain which mark is to be considered the student's item reponse. The microprocessor can score the test and in some scanners cause the score to be printed on the answer sheet. It is also possible to compute summary statistics as the sheets are processed. The microprocessors are also used to control the communications process and coordinate data with a computer. The inclusion of the microprocessor provides the desk top scanner with features and capabilities previously found only in the expensive high

capacity scoring systems. Yet, the overall cost of the desk top scanner has been quite stable for a number of years.

The linking of the desk top scanner to a microcomputer results in a small scale test storing and reporting system. Microcomputers such as the APPLE II, IBM PC, or TRS-80 have more than enough computing power to score a test, compute summary statistics, perform item analyses, and print a variety of reports. Thus, for about $5,000 a school can have an automated test scoring system that can process a very respectable number of tests per day. Recently, National Computer Systems has begun to market just such a system consisting of an IBM PC and an NCS 3000 desk top scanner. The Nippon Electric Company has been marketing a microcomputer-based test scoring system for a number of years. The system includes a mark sense card reader, a microcomputer, a typewriter printer, and an S-P keyboard. The latter is a special device used to enter item reponse data manually for a specific type of item analysis. Systems of this type should find widespread acceptance in the schools.

All of the test processing systems described to this point have used answer sheets upon which the students mark their item response choices. The Test Input Device (TID) (Syscon, 1983) is designed to eliminate the answer sheet and the OMR from the test processing sequence. The TID is about the size of an electronic calculator and has a similar keyboard and display. Internally it contains a microprocessor and about 20,000 bytes of computer memory. A probe on the device allows it to be connected to a computer via a "black box". Initially, the computer is used to prestore in the TID information about the test such as the test identification and the number of

items. In use, the student employs the keyboard to enter his ID number, the test ID, and then his response choice for each item. A feature of the device allows the student to review his choices, change them, and otherwise edit them via the keyboard and the display. When the student is done, he simply inserts the probe into the "black box", presses a button, and all the data are transmitted to the computer, where the usual programs score the test and report the results. The U.S. Navy is currently using the TID and appears to be pleased with its use as a test answering vehicle. The technology contained in the TID is a spinoff of military data collection, and the device itself is similar to those used to audit inventories in, say, a grocery store. Given the high cost of answer sheets, use of the TID could be quite economical in many testing situations.

A significant facet of any large scale testing program is the creation and maintenance of the item pool. This has led to the use of computers, and a field known as item banking (Wood & Skurnik, 1969; Lippey, 1974) has arisen in testing programs. The basic idea is to store the actual test items in mass storage and provide functions that allow one to inspect items, edit them in various ways, and select them for inclusion in an instrument. Since an item pool exists to support the test construction process, most computer-based systems merge the item banking and test construction process into a single software package (Lippey, 1974).

A decade ago, I developed an item banking/test construction program based upon an item response theory approach that still defines the state of the art (Baker, 1972). This program maintained an item pool, kept historical records of the item and the test statisics, and

was integrated with a test scoring program. When a test was scored, all the historical data such as item and test statistics were automatically updated in the files. This computer program was implemented on a large scale computer and designed to be used by a relatively sophisticated test constructor. As one would expect, the item banking/test construction process has been implemented on a microcomputer (ATA, 1981). This system employs a TRS-80 with two floppy disks, and the computer program is called TEST BANK. The program can store about 300 items and the user can select items using a limited number of keys. Despite its modest capacity, this system represents a significant contribution to testing technology. As the data storage capabilities of microcomputers grow, so will the size of the item banks that they can maintain. More importantly, such systems make item banking/test construction accessible to a very wide range of users.

Computer administered tests have been employed for many years, primarily via time sharing terminals from a large computer. Such testing has always been expensive and limited by the small number of terminals available on any one computer. The microcomputer has opened up new and interesting possibilities for on-line testing. It is quite possible to store test items on a floppy disk and have the microcomputer administer the test and record the student item reponses on a disk. Upon completion of a testing session, a student could be given immediate feedback by the computer as to the results and their interpretation in either a norm referenced or criterion referenced sense. In addition, the graphic capability of the micro can be used for figures and diagrams appearing within individual items. Voice

output devices such as the "type and talk" (Votrax, 1983) can be used to present questions and instructions verbally. Clearly, the microcomputer capabilities such as graphics, verbal output, data storage, and man-machine interaction can be forged into a very powerful and dynamic testing vehicle. The present state of the art offers all of these capabilities at a modest hardware cost. Microcomputer systems for administering standard psychological tests such as the MMPI are currently used in a number of settings. Such testing is also offered as a commercial service by a number of consulting psychologists.

One form of computer administered test is the tailored or adaptive test first proposed by Lord (1970) and investigated extensively by Weiss (1975). Under the adaptive testing procedure, a series of items are administered via a set of rules that select items appropriate to the examinee's ability. The available schemes narrow in on a student's ability level rather quickly, and each student gets a unique test. Such adaptive testing rests upon item response theory and is an excellent example of the application of the theory to practical testing. Given a precalibrated item pool, the actual item selection, test administration, and scoring procedures of the adaptive testing can be implemented easily upon a microcomputer. Such adaptive testing is particularly appropriate for schools using individualized instruction where students are tested on an as-needed basis rather than in groups at predetermined times. Further it does not require personnel to administer the test. These and other advantages have led to the creation of microcomputer systems devoted specifically to adaptive testing. Vale and his associates in Minnesota are currently developing such a system for the military.

Computer aided instruction (CAI) and its derivatives, such as computer based instruction (CBI), have had a long and rather tortuous history. In the mid 1960's it was treated as the savior of education and considerable resources were devoted to the area without a great deal of success. By the mid 1970's the CAI furor had subsided and only a minimal interest existed. However, the emergence of the microcomputer, and its widespread penetration into the home and the schools, has given CAI/CBI a dramatic new lease on life, and it has risen Phoenix-like from the ashes. Numerous sources report that instructionally-oriented software for microcomputers is one of the major growth industries in the 1980's. Many new software houses have been established to produce such software, and old-line publishing houses are getting deeply invoved in the selling of instructional computer programs.

Inspection of this new generation of instructional software quickly reveals that it is nearly always devoid of any testing or evaluation component. Drill and practice programs will tell the student how many problems they got correct, but most other instructional computer programs do not. The general approach seems to be one of letting the student interact with the computer in various ways; when the student reaches the end of the program paradigm, the computer says goodbye and the student walks away. In addition, there are no records kept for the teacher to use in evaluating student progress. While such software may in fact be instructive, it does not integrate well into the overall educational process. The lack of a testing or evaluation component in such software is particularly

disconcerting.

At this point, I would like to briefly summarize the state of the art in testing technology.

1. The desk top scanner and microcomputer combination provides the technological base for a cost-effective test scoring and reporting system. As a result, most organizations and schools can now afford the test processing associated with small to moderate scale testing programs. Such a capability allows schools much greater flexibility and control over their testing programs. The state of the art in high capacity test scoring and reporting is at a sophisticated level. Because of this, the data processing aspects of nationwide or other large scale testing programs can be performed efficiently and at a reasonable cost per student.
2. Due to microcomputers, item banking at the local level has become feasible. However, the practical size of the item pool is still somewhat limited by the available storage devices.
3. The existence of a dedicated hardware/software system for delivery and administration of tests is an important advance. In particular, the development of microcomputer based adaptive testing is an important increment to the state of the art in testing.
4. All indications suggest that the computer as an instructional device is finally going to make sufficient penetration into local schools to be considered a viable technology. However, existing instructional software rarely includes any provision for testing or evaluation.

FUTURE TRENDS

Test Theory Trends

For those of us who work in psychometric theory, it has been clear for many years that classical test theory has run its course. Classical test theory grew out of Spearman's work on intelligence during the 1920's, and was the theoretical underpinning of the past half century of testing practice. However, a close look at classical test theory shows that other than some work on generalizability (summarized by Brennan, 1983), there has been little if any extension or elaboration of this theory in recent years. In addition, existing testing procedures and practices have fully exploited the capabilities of the theory. The theory is a mature one and its future growth does not seem likely. In sharp contrast, item response theory (IRT) is alive, dynamic, and growing rapidly. A considerable body of new theory is in place due to the efforts of Lawley (1943), Lord (1952), Bock (1972), Samejima (1972), Wright (1967), Wright and Stone (1979) and others. This theory is vastly superior to classical test theory in its conceptualization, since it is based upon the item rather than a test score, and has considerable potential for further growth.

At the present time, IRT is rapidly making the transition from a pure theory to one that is widely used in practice (Lord, 1970). In addition, it has provided analysis techniques for estimating parameters of items having graded or nominal response that classical theory could not handle. I believe that the future will see an acceleration of the transition from a testing practice based upon classical test theory to one based upon item response theory. Because

of its greater mathematical sophistication, the application of item response theory is going to depend heavily upon technology.

Technological Trends

The microcomputer revolution is not even a decade old, and the rate of change still appears to be accelerating. First we had microcomputers which were 8 bit machines; next we had the 16 bit machine; today we have 16 bit machines with 32 bit internal registers. The net result is that today's microcomputers are rapidly exceeding the capabilities of the medium scale computers of a few years ago, at a fraction of the hardware cost.

Although the hardware currently available has considerable power, that which is just around the corner is even more startling. For example, computers based upon the Intel 432 chip are just now beginning to emerge from the R and D shops. This family of computer chips allows one to create 32 bit microcomputers that are as powerful as today's large scale computers. Yet they are physically no bigger than the familiar personal computers. In a very real sense, this will be like having your university's computer center sitting on your desk. Thus, an increasingly large amount of raw computer power is becoming available to use within the context of testing programs and related applications. The only factor to dampen our enthusiasm for this class of machines is that the software to make use of it will cost many times that of the actual hardware. One may pay $15,000 for the computer and then $50,000 for the operating system, language compilers, mathematics libraries, statistical packages, etc., that are necessary to exploit the hardware.

The second area of very rapid technological advance is that of mass storage. The basic trend is toward ever larger storage capacity at a relatively decreasing cost per unit of storage. For example, most microcomputers have used a "floppy disk" that can store about 144,000 characters on a disk. With today's technology, this unit can be replaced with a unit of the same physical size that stores several million characters. Slightly larger yet modestly priced units can store from 10 to 40 million characters per disk.

It should be noted that a major limiting factor in a number of areas of testing such as item banking, test construction, and adaptive testing has been the lack of low cost mass storage. The increasing availability of low cost mass storage promises to have a significant impact on testing procedures dependent upon an item pool. Given the availability of such mass storage, the technical limitations on our ability to create and maintain such item pools are rapidly disappearing. It will also increase the feasibility of on-line adaptive testing as an item pool, for longer sequences of items can be accessible to the algorithms that select items for administration. From a practical application in testing point of view, the increase in low cost mass storage will probably have a greater impact upon the field than the increase in raw computing power. This is because most of our testing applications are heavily data based oriented and only marginally number crunching oriented.

Perhaps the ultimate in mass storage are the new optical storage devices. Some of these are "write once, then read only" devices. Once the data are stored they can never be changed. Such devices have considerable promise for archival data, such as obsolete item pools,

but could not be used for active item pools. However, optical mass storage devices that can be used like ordinary disk storage are beginning to appear and offer exciting possibilities in very large scale data storage. It will be a few years before this technology gets within the price range most testing organizations can afford.

A related memory device is the video disk which allows one to store full video screens and play them back. This opens up the door to dynamic presentation of test items such as the re-enactment of a historic event or the recording of a physical process. At the present time, the cost of material development and creation of a master disk is very high. However, copies of the disk are not too expensive. Thus, it would be possible to use this technology for testing, given the proper equipment. A few systems combining a microcomputer, a video disk player, and a color TV set are available commercially. But I am unaware of their use within the context of testing.

A third technological trend of interest is that in optical mark reading equipment. The introduction of the desk top scanner in 1974 has provided diverse organizations with a reasonably priced means of scanning answer sheets. There has been a trend toward greater sophistication within such scanners, but the cost has remained near the $3,000 per unit level. What is really needed is a low cost, say $600, desk top optical mark reader. Such a device is not as complicated as a printer, and the state of the art of optical marker reading is adequate; hence, to create and produce such a piece of equipment is not a major problem. However, I suspect that when such a scanner appears it will carry a Japanese nameplate. The availability of a really low cost scanner will put automated scoring and reporting

systems at the classroom level.

An age-old problem in the production of tests is including pictures and diagrams within the test item. Traditionally this has been done manually via cut and paste procedures. Fortunately, there are a wide range of commercial and military applications that face the same problem. For example, technical manuals are widely used and include both text and engineering diagrams that must be revised frequently to keep the manual current. To meet this need, equipment has been developed that can scan graphics material, convert it into a computer representation, and save it on a mass storage device. It is then a simple matter to merge these digital representations of the graphics with textual material and produce both the graphics and text on the screen of a video display terminal, or print them on paper.

It should be noted that a variation of this can be done with a personal computer using a GRAPPLER II board and an EPSON MX-80 printer with GRAFLEX chips. One can program the desired diagrams using high resolution graphics, and store the binary file on the floppy disk. The textual part of the question can be programmed and stored. To reproduce a hard copy of the item, the graphics information is read into memory and displayed on the VDT screen. With a single command, the diagram is reproduced by the printer. Then the text portion can be printed. This is not as nice as simply scanning the diagram, but it does provide a significant test item creation capability. Numerous computer programs, such as the Graphics Magician (Pelczarski, Lubar, & Jochumson, 1982) are available to facilitate the creation of the graphics part of the item.

One of the really active areas of technological development is that of computer networks. The goal here is to create a communication network through which many different computers "talk" to each other. Where all the computers are in a reasonable proximity to each other, these schemes are called "local area networks". The driving force behind such networks is the automation of the office and the need for large corporations and/or government agencies to exchange data on a computer-to-computer basis. Perhaps the best known of the nationwide networks is the ARPA net that interconnects many universities and governmental laboratories. A major thrust is to interconnect microcomputers, and this can be done using commercial networks such as APPLENET.

Computer networks have a direct application to a wide range of testing procedures. For example, in the Netherlands an effort is underway to create a nationwide test processing network. Each school will be equipped with a desk top scanner, a microcomputer with disk drives, and a printer. When a testing program is conducted, the answer sheets are scanned locally, and the results stored on disks and then transmitted to a central computer. Upon scoring, item analysis, norming, etc., the test reports for the school and the individual students are to be transmitted back to the schools for printing via the microcomputer. At the same time, test results aggregated by school districts and other larger units are available at the central computer. While similar systems have been built in the past via the time-sharing capabilities of large computers, the microcomputer-based network offers much greater flexibility and ease of use.

Another application of computer networks is in on-line testing. It is now technically feasible for a central computer with some mass storage to store many different item pools. A student can sign on to a microcomputer in the network, request to take a test, and have his request validated. The central computer will then transmit a test (possibly unique to the student) to the microcomputer, which then administers the items, scores the test, and tells the student the results. Upon completion, the item responses and other data are transmitted back to the central computer for aggregation and storage. Again, such a scheme has been possible in the past, but the microcomputer and local area networks place this within the realm of the readily achievable.

The final technological trend of interest to testing practice is in the area of software. Until quite recently, most applications implemented on a computer existed as separate computer programs with their own set of procedures and purposes. The disadvantage of this approach was that the user had to learn the procedures for each application in total isolation from all the other applications. Each had its own set of control functions, unique features that were tailored to the problem, and no commonality of logic. The result was a rather large learning period for a person who needed to use several different applications. About ten years ago, the trend was to place a common data base underneath these applications. Even though each application proceeded independently, they nonetheless used the common set of data. Within the past 2-3 years, efforts were initiated to integrate a number of seemingly disparate applications into a single

coherent package. The first available integrated package was implemented on the APPLE LISA computer system. The software provides an integration of word processing, spread sheet calculations, business graphics, list maintenance, PERT charting and computer terminal emulation. All of the data employed by a user of the system can be passed easily from one application to another via a simple user-friendly set of procedures.

A number of other such integrated systems are following on LISA's heels, and they will be commonplace in the near future. Although these integrated systems have been designed for the business environment, the basic approach and techniques are directly applicable to the testing environment. It is currently feasible to create a hardware/software system for a microcomputer that could integrate item writing, item banking, test construction, on-line administration of tests, test scoring, item analysis, and reporting of test results. Once the appropriate software tools are made available, this might even be done within the context of an existing system such as LISA. In any event, it could be done if one were to devote some people and resources to the task.

Instructional Trends

The widespread penetration of the microcomputer into the schools is beginning to result in change in instructional approaches. At the college level, a large number of textbooks are being accompanied by a floppy disk which contains computer programs to be used in conjunction with the text. These progams range from simple exercises to sophisticated simulations of complex processes. In the physical

sciences, many of these programs enable students to explore topics that would be prohibitively expensive to implement in a laboratory setting. A similar pattern is beginning to develop at the secondary and elementary level, where computer software is being used to supplement existing texts and to provide enrichment. As mentioned in an earlier section, textbook publishers are moving quite rapidly to establish a market share in what is being called "electronic publishing". This activity suggests that the publishers see an underlying trend in which they must participate to ensure future business.

At the present time, the coordination between the textbooks and the computer software varies from very loose to a reasonably good level. Much of the software has been collected from a variety of sources and has been pooled under a commmon title rather than having been created specifically for the text. However, the longer term trend is toward a closer linkage between the material and approaches taken in the text and those in the computer software. At some point in time, instructional designers, curriculum specialists, psychometricians, textbook writers, and software developers are going to work as a team to jointly develop curriculum, instructional software, evaluation procedures, and instructional management systems, all within a common frame of reference.

It should be noted that the result will not be classical computer-aided instruction or computer-based instruction. Rather, the result will be textbooks written to take advantage of the educational leverage the computer can provide. Under this approach, an instruc-

tional decision is made that a specific topic can be handled better via the computer than by the textbooks or some other vehicle. Only then would a computer program be written and its use integrated into the instructional flow of the text. In many places, it would be clear that the text or other materials would be more appropriate. The net result is going to be a mixed bag of conventional and computer-based procedures, all of which contribute to the instructional process defined by the text. In addition, the computer will maintain records of information needed by the teacher to effectively monitor and manage student progress through both the conventional and computer-based parts of the course.

The issue of present concern is where testing fits into such a highly integrated system. Conventional testing will be used to measure student progress in the broader sense. However, within the context of the computer-based aspects of the curriculum, a significant change will occur. The old "pretest, instruct, posttest, remediate" paradigm employed so widely in computer-aided instruction will be abandoned. This paradigm has been with us since the days of programmed instruction (Coulson, 1961), and is badly timeworn. The basic problem with the paradigm, within the context of computer usage, is that the student spends too much time answering multiple choice questions rather than using the computer in an optimal fashion. Present-day microcomputers have sufficient power to enable the student to use the computer as an exploratory tool as well as to implement highly dynamic modes of instruction. Time spent responding to multiple-choice items detracts from the student's productive use of the computer.

The testing alternative that I see developing is what could be called "non-intrusive" testing. Under this approach, a student using an instructional software package would never be administered a formal test within the context of using the computer. Instead of recognizable tests and test items, instructionally relevant data would be collected as the student interacts dynamically with the instructional software. A variety of information, such as the sequence in which capabilities of the software are employed by the student, the rate at which a student moves from one level to the next of Bloom's taxonomy, and the strategies employed by the student to reach instructional goals, can be collected during the computer session. Given this data, evaluation routines embedded within the instructional software can ascertain the student's instructional status. An excellent example of this type of evaluation is the model developed by Brown and Burton (1978) to identify "bugs" in a student's learning procedures. When the desired level of understanding has been reached, the computer simply tells the student he knows the material and should move on. From the student's point of view, a test was never taken; however, from an evaluative point of view, the student has been continuously evaluated, and teachers have at their disposal a wide range of evaluation data collected by the computer.

Such non-intrusive testing clearly makes better use of the student's time as well as the computer resources. While few examples of non-intrusive testing procedures exist at present, there are some antecedents. The standardized grade score used by Suppes and Morningstar (1972) was computed dynamically while a student did drill and practice exercises. After each problem, this score was recomputed

and used to select the next problem to be used. The diagnosis and prescription procedures implemented under CMI have similar characteristics even though they are based primarily on test scores (Baker, 1978). The analysis portions of non-intrusive testing have much in common with the current efforts in artificial intelligence and in particular with "expert systems". The latter are computer hardware/software systems that embed the decision-making heuristics of experts within the software, which is then used by less skilled persons to analyze situations and draw conclusions. The immediate example is medical diagnosis, but educational diagnosis could just as well have been the area of interest implemented via an expert system.

Summary of Trends

Let me briefly recall the future directions as I see them.

1. Clearly, classical test theory has reached its upper limit of development and will be replaced in practice by item response theory.
2. Microprocessor-based technology will continue to move at a rapid pace. This will make it easier to automate existing practice. It will also provide the basis for developing specialized hardware/software systems for use in the field of testing.
3. There will be an increased emphasis upon a coordinated approach to instructional design that exploits the microcomputer as an educational vehicle. Non-intrusive testing could be a significant part of this approach.

BARRIERS AND PROBLEMS

Conventional Testing

Let me mention several old problems before describing a new one. Modern technology, in the form of desk top scanners, microcomputers, and high-capacity test scoring equipment, makes the automation of test scoring and reporting easy and cost-effective. However, there are limits to the amount of testing that the schools can absorb. We have seen several swings of the testing pendulum in the past few decades, and regardless of the level of testing, the technology for processing the results can handle the workload.

Modern technology has also provided the means for maintaining very large item pools, and providing automated or semi-automated item selection from these pools. It is not uncommon to have pools of up to 25,000 items stored via a computer system. The quality of these item pools is yet another matter. Whenever such a large pool exists, it is usually the result of having many people in many different settings write items and enter them into the pool. In such circumstances, it is extremely difficult to maintain the instructional focus of the items as well as to ensure technical quality. For example, Brenner (1981) reported, in the case of a pharmacology item pool of 25,000 items, that only 6,500 items were retained after scrutiny by a review panel of subject matter specialists. Most large item pools would exhibit similar shrinkage upon close inspection. The underlying message here is that technology can make a process easy to implement, but it does not ensure the quality of the product.

Technology has contributed indirectly to a new testing problem. In an earlier era, the development of tests was a rather academic process. Typically, a scholar became interested in a subject matter area and constructed a test. The instrument was refined in the course of a few school years. Graduate students would further explore the instrument through their thesis research. After a few years, the instrument was reasonably well developed, and some research existed that described the reliability, validity, and interpretation of the instrument. In some cases, the scholar printed and distributed the test in the marketplace. In other cases, the test would be taken over by a commercial testing organization and marketed. Even when instruments are developed within a testing company, the basic paradigm holds with a somewhat different cast of workers. On the whole, test development was a rather gentlemanly pursuit that was only sporadically impinged upon by forces outside of the educational establishment.

Beginning about 20 years ago, in the context of employment screening instruments, political forces intruded into testing. At the present time, two politically motivated events have occurred that have a major impact upon testing. First, the competency testing of the 1920's and 30's has been resurrected from the grave and given new life. Second, in some states, item disclosure laws have been passed which give examinees access to items in the tests they have taken. I don't intend to argue the merits of these events, but the process by which they came about and their impact upon testing bear some examination. The politicians who pushed these two ideas did so with

little or no understanding of the nature of test development, the technical issues, or the long-term ramifications for educational measurement. The net result is that measurement specialists have been thrust into a situation for which they are ill prepared.

In the case of competency testing, there is a demand for immediate large-scale testing on a wide range of subject matter, with little or no underlying test development. The item disclosure rules result in tests that have been carefully developed over many years being put in the public domain. If the test is to be retained, it forces those reponsible for the test into a high-speed, iterative item development process to replace compromised items. Although there are many forces contributing to an increased role of the politicians in the measurement arena, one of the culprits is technology. The general public's naivety about how computers work carries over into the political arena. What happens is that the politicians hear that tests are scored by computers, that item pools are maintained on the computer, and that tests can be printed via computer. The conclusion reached is that if so much of testing is automated, it must be a simple matter to establish a testing program; you just let the computer do it! I suspect that in the long term, the ability of measurement specialists to cope with these external inputs will depend upon that same technology that helped get them in trouble.

Technological Barriers and Problems

Technology itself is also a barrier and a problem. Because of the rapid pace of technological development, the hardware in particular advances faster than we are able to absorb it into the

daily world of testing. Taking advantage of new technology, such as optical storage, involves major levels of effort. It takes time and money to explore what testing uses can be made of the technology, and the start-up costs are independent of the eventual use of the technology in the schools. In addition, making the transition from a feasible use of technology to one that is widely used in the schools involves a very high level of effort and cost. Even when such technology does reach the schools, it requires resources which are scarce. Hardware must be maintained. It becomes technologically obsolescent rather quickly, and the trade-in value of an old piece of hardware is nil. Thus schools have to seriously consider the cost-benefits ratio when introducing technology of any kind. In particular, any testing-related technology must have a favorable cost-benefit ratio in order for it to be widely adopted by the schools.

While the creation of hardware/software systems that integrate item writing, item banking, test construction, on-line or adaptive testing, graphical capabilities, automated test scoring, and reporting will occur, it is a significant development task. For example, the LISA system is reported to have cost $50 million to develop. An integrated testing system is at an equal level of technical complexity, yet the potential market for such a system is very small compared to that for a work station such as LISA. As a result, the development of such systems is going to be an evolutionary process based upon available technology rather than a sudden quantum step. The components are all there; it is their integration into a flexible,

powerful system with sufficient generality that costs time and money.

The Manpower Barrier

The current testing milieu is one in which testing is not as static as it once was. The field is more dynamic, more dependent upon technology, and increasingly under greater scrutiny More importantly, the context within which testing is employed is becoming increasingly unstable. Because of this, it is increasingly difficult to construct items and to refine and polish them in a volume necessary to meet the need. One consequence of this will be a lowering of the quality of the tests due to insufficient development. The conclusion is that maintenance of quality both in terms of content and technical characteristics requires more trained personnel than are currently devoted to test development. In addition, if testing is going to exploit technology, people are needed who can work within both measurement and hardware/software. This is particularly crucial if adequate integrated systems are to be developed to support testing. The trained manpower barrier to future development of testing is not an obvious one, due to the diffusion of such manpower across a wide spectrum of levels in government and educational institutions. Yet it does exist.

IMPLICATION AND POLICY RECOMMENDATIONS

Education

The rapid transition from classical test theory to IRT has many implications for the use of tests in the schools (specifically, the reporting of test results in an ability metric that will do much to improve the interpretation of test results). IRT allows the use of

new item types as well as new testing procedures such as adaptive testing. The policy recommendation is twofold. First, schools and other responsible agencies should foster the switch in the psychometric underpinnings of testing practice. Second, vehicles need to be put in place to fully explore the ramifications of item response theory for the day-to-day practice of testing.

The increasing use of microcomputers in the classroom opens the door to the use of "non-intrusive" testing procedures. However, this idea is not well formed at present, and considerable research needs to be done to determine the underlying principles of such testing. Without such research, each application of the approach is a special case, and it will be difficult to determine if the basic concept is viable. Once the basic framework of non-intrusive testing is established, the incorporation of such testing into modern computer-supported curriculum will be much easier.

Conventional testing as we now know it is going to be with the schools for a long time. However, the demands upon the schools for both externally and internally imposed testing will increase with time. The major implication is that there will in all probability be a continuing shortage of personnel trained in measurement and its related technology to develop such tests.

One of the clear outcomes of CSE's ongoing study of test use is that the majority of classroom testing employs teacher-made tests. Yet, classroom teachers are provided with very little assistance in the preparation of such tests. Thus, efforts should be initiated to develop a microcomputer-based test development system for use by classroom teachers. Such a system is within the state of the art, and

its availability could have a significant impact upon testing in the schools.

Technology

There currently is a significant lag between the introduction of a new level of technology and its application to the field of testing. What is needed are vehicles so that this technology can be employed quickly and its advantages/disadvantages for use in testing can be determined. Such early evaluation allows one to both discover viable uses of such technology and to enable others to avoid nonproductive uses of the technology. Let me briefly describe some examples of areas of technology where pilot projects would be valuable.

The availability of low-cost mass storage for microcomputers has major implications for item writing, item banking, test construction, and on-line adaptive testing as well as for automation of test scoring and reporting. We need to look into what can be accomplished using these storage devices. In particular, their role in the development of tests needs to be explored.

Video disk technology opens many possibilities in both conventional and non-conventional testing procedures. It makes possible test items that involve dynamic presentations as well as active examinee participation in the evaluation process. The Achilles heel of this technology is the enormous material preparation time. The role of video disk technology in testing, as well as vehicles for minimizing the material preparation time, needs to be examined.

The leading edge of the application of computer technology currently deals with computer networks. The hardware and software is available to construct a wide variety of networks. Such networks have many implications for testing. These involve down-loading of tests to local sites from central sites, aggregation of test results across widely distrbuted sites, and flexible mixes of conventional on-line and adaptive testing. The interesting feature of this work is that it is focused upon allowing microcomputers to be networked. The ramifications of networking for testing need to be investigated.

Pilot projects in these and other areas can be conducted in a variety of settings and are within the capabilities of a range of educational institutions. The results from such pilot projects would do much to set the tone for the improvement of testing via technology.

SUMMARY

The intent of this paper was to provide an overview of the symbiotic relationship between testing and technology. This relationship has been developing since the earliest days of the testing movement. Despite the age of this relationship, it has not gone awry. One of the major factors in its continuity is that the cost of high technology has been reduced to the point where it is accessible to most of those with an interest in testing. Because of this, one is as likely to see sophisticated research and development projects dealing with testing at the local school district level as in professional educational innovation organizations. As a result, these are rather exciting times in the field of educational measurement. Hopefully, one outcome of this paper will be the addition of further excitement to the relationship of testing and technology.

REFERENCES

Baker, F.B. Automation of test scanning, reporting and analysis. In R.L. Thorndike (Ed.), Educational Measurement Washington DC: American Council in Education, 1971.

Baker, F.B. A conversational item banking and test construction system. Proceedings of the Fall Joint Computer Conference, AFIPS, 1972, 41, 661-667

Baker, F.B. (1978). Computer managed instruction: Theory and practice. Englewood Cliffs: Educational Technology Publications, 1978.

Baker, F.B. Automated scoring systems. In T. Husen & T.N. Postlethwaite (Eds.), International encyclopedia of education: Research and studies. Oxford: Pergamon Press, 1983.

Bock, R.D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories Psykometrika, 1972, 37, 29-51.

Brennan, R.L. Elements of generalizability theory. Iowa City: American College Testing Publications, 1983.

Brenner, L.P. On-line item banking in health sciences education. Journal of Educational Technology Systems, 1980-81, 9, 213-227.

Brown, J.S., & Burton, R.R. Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 1978, 2, 71-109.

Coulsen, J.E. (Ed.). Programmed learning and computer based instruction. New York: Wiley, 1961.

Lawley, D.N. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 61A, 273-287.

Lippey, G. (Ed.). Computer-assisted test construction. Englewood Cliffs, NJ: Educational Testing Publications, 1974.

Lord, F.M. A theory of test scores. (Psychometric Monographs No. 7), 1952.

Lord, F.M. Some test theory for tailored testing. In R. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1970.

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Marco, G.L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1980, 14, 139-160.

Pelczarski, M., Lubar, D., & Jochumson, C. The graphics magician. Geneva, IL: Penguin Software, 1982

Samejima, F. A general model for free response. [Monograph 18]. Psychometrika, 1972, 37, Part 2, 1-68.

Suppes, P.C., & Morningstar, M Computer-assisted instruction at Stanford, 1966-1968: Data, models, and evaluation of the arithmetic programs. New York: Academic Press, 1972.

The Test Bank. San Diego: Advanced Technology Applications, 1981.

The Test Input Device. Washington, DC: Syscon Corp, 1983.

Type 'N Talk User's Manual. Vortrax, 1981.

Weiss, D.J. (Ed.). Computerized adaptive trait measurement: Problems and perspectives. (Research Report 75-5) University of Minnesota, Psychometric Methods Program, Department of Psychology, 1975.

Wood, R. & Shurnik, L.S. Item banking. London: National Foundation for Educational Research in England and Wales, 1969.

Wright, B.D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems (pp. 85-101). Princeton NJ: Educational Testing Service, 1968.

Wright, B.D., & Stone, M.H. Best test design. Chicago: MESA Press, 1979.