DOCUMENT RESUME

ED 266 168 TM 860 112

AUTHOR	Estes, Gary D., Ed.
TITLE	Examples of Item Banks to Support Local Test Development: Two Case Studies With Reactions.
INSTITUTION	Northwest Regional Educational Lab., Portland, OR. Assessment and Evaluation Program.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Nov 85
CONTRACT	400-83-0005-P-15
NOTE	75p.
PUB TYPE	Reports - Descriptive (141) -- Collected Works - General (020)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Adaptive Testing; Case Studies; *Computer Assisted Testing; Costs; Elementary Secondary Education; Flow Charts; *Item Banks; Latent Trait Theory; Microcomputers; Minimum Competency Testing; *Teacher Made Tests; *Test Construction; Test Format; Testing Programs; Test Items; Test Use
IDENTIFIERS	*Portland School District OR; *Wisconsin Item Bank

ABSTRACT

This report and compilation of papers summarizes information collected by an Assessment Development and Use Project, initiated by the Northwest Regional Educational Laboratory (NWREL) to assist test development efforts by state and local agencies. Specific item banking applications are reported in two case studies, selected because they represent item bank efforts in which agencies other than the ones that developed the item bank have been able to develop tests from these item banks. The first item bank described is the Wisconsin Item Bank, developed by the Wisconsin State Department of Public Instruction to be used by school districts participating in Wisconsin's voluntary competency-based testing program. The design, history, problems, and operational details of the item bank are discussed in detail. The second report describes an item bank developed by the Portland Public Schools. This item bank has supported Portland's local testing needs and has been shared with several other districts. Discussed are the history and purposes for development, current uses, problems and issues in development, management, and use, and ways the technology may be shared. The final section of the paper considers issues related to how well item banks work and how well they fulfill their potential. (LMO)

TM

ED266168

Examples of Item Banks to Support Local Test Development:

Two Case Studies With Reactions[1]

Gary D. Estes
(Editor)

Evaluation and Assessment
Northwest Regional Educational Laboratory
300 S.W. Sixth Avenue
Portland, Oregon 97204

November, 1985

TM 860 112

[1] This set of papers is based upon work performed pursuant to Contract 400-83-0005 P-15 of the National Institute of Education. It does not, however, necessarily reflect the views of that agency.

TABLE OF CONTENTS

Page

I. INTRODUCTION 1

Gary Estes
Northwest Regional Educational Laboratory

II. THE WISCONSIN ITEM BANK 5

Nancy W. Burke
B. Darwin Kaufman
Norman Webb
Wisconsin Superintendent of Public Instruction

III. A REGIONAL AND LOCAL ITEM RESPONSE THEORY BASED TEST ITEM BANK SYSTEM 28

Walter Hathaway
Ron Houser
Gage Kingsbury
Portland Public Schools

IV. HOW CLOSE ARE WE TO REALIZING THE POTENTIAL OF ITEM BANKS 59

Gary D. Estes
Northwest Regional Educational Laboratory

I. INTRODUCTION

Gary D. Estes
Northwest Regional Educational Laboratory

The role of assessment in local school districts and state education agencies continues to be an important concern. Tests have long been a primary benchmark by which effective schools or education are measured. Decreases or increases in test scores are often the most public measures of school and district effectiveness. In addition to these accountability stimuli, the literature and information on effective school practices support the roles of monitoring and assessment in promoting high student achievement. Thus, both improvement and accountability objectives have led to high interest in local use and management of test information.

One result of this empahsis has been to stimulate districts and states to develop their own test items or resources which could be used to assess local goals. Another result of increasing emphasis on testing is concerns about budgeting and testing cost and ways to maximize information while minimize costs. Advances in technology and item banking have changed the capability for developing and managing tests and other information.

In response to this need, the Northwest Regional Educational Laboratory (NWREL) initiated an Assessment Development and Use Project to assist test development efforts by state and local agencies. This report summarizes much of the information that has been collected. A brief overview of work that led to this compilation of papers will serve as general information in the area of item banking and technology to support local test development efforts.

Item Bank Surveys

NWREL surveyed every state for information on item banking and local test development activities within each state. These phone call interviews resulted in a mail survey to 125 agencies that had item banks or information that might be shared with others. Of the 125 mail surveys, 54 were returned and information on 41 item banks were included in A Guide to Item Banking in Education (Second Edition), (Estes & Arter, 1984). The Guide provides descriptions of item banks around the country that have information that might be shared with other agencies. Sources of item banks include state departments, school districts and several test publishers.

Technology has also played an increasing role supporting state and local test development efforts. The Assessment and Development Use Project has also reviewed several microcomputer programs that can be used to support local test development efforts. These are summarized in Deck and Estes (1984), and Deck, Nickel and Estes (1985).

This report and compilation of papers builds upon the earlier project work by providing reports on specific item banking applications. These case studies were selected because they represent item bank efforts in which agencies other than the ones that developed the item bank have been able to develop tests from these item banks. The first item bank described is in the State of Wisconsin. The state department has developed an item bank from which local districts can develop tests matched to local objectives. The Portland Public Schools reports on an item bank developed over several years. This has supported not only Portland's local testing needs but has also been shared with several districts in the Northwest and throughout the United States through the Northwest Evaluation Association which is a consortium of districts. Reports on these item banks include the history and purpose for development, the ways in which technology and measurement methodology have been used in developing and managing the item banks. They also offer ideas on how information can be shared and next steps needed to further the development and usefulness of item banking for state and local test development. A review and reactions to these item banks reflects on the degree to which the advantages of item banking is being realized and offers some suggestions and cautions to others interested in undertaking item banking as an approach to support local assessment efforts.

Finally, Arter and Estes (1985) have produced a Handbook for Item Banking that is designed to guide a district or staff through a decision process that leads to deciding whether someone else's item bank, a locally developed item bank or other systems, e.g., commerical tests, will best serve local testing purposes.

References

Arter, J., & Estes, G. (1985). Handbook: Item Banking for Local Test Development. Portland, Oregon: Northwest Regional Educational Laboratory.

Deck, D., & Estes, G. (1984). Microcomputer Software Support for Item Banking. Portland, Oregon: Northwest Regional Educational Laboratory.

Deck, D., Nickel P., & Estes, G. (1985). Reviews of Microcomputer Item Banking Software. Portland, Oregon: Northwest Regional Educational Laboratory.

Estes, G., & Arter, J. (1984). A Guide to Item Banking in Education. Second Edition. Portland, Oregon: Northwest Regional Educational Laboratory.

II.

The Wisconsin Item Bank

The Wisconsin Item Bank is a collection of over 10,000 test items for grades 3 through 12 in reading, mathematics, and language arts. These items are stored in a computerized system. This item banking system, developed and operated by the Wisconsin Department of Public Instruction, is used by school districts participating in Wisconsin's voluntary competency-based testing program. The Bank, developed under a 1981 mandate from the Wisconsin Legislature, began operation in July 1, 1984. During its first year of operation, the Wisconsin Item Bank provided customized tests to 35 of the 140 districts participating in competency testing.

This paper describes the history and development of the Wisconsin Item Bank as well as what has been learned about developing item banks in general. Decision makers, who are considering developing or using an item bank, are the intended audience. While the system was designed to support local school districts in a competency testing program, many of the steps and problems faced are common to the development of any item bank system at any level: national, state, regional, or local. An effort has been made to state clearly the assumptions and constraints that were placed on the development of this system and to separate the problems specific to the Wisconsin Item Bank from those problems which can be generalized to the development of any item bank.

The authors of this paper each had different roles in the development and the operation of the Wisconsin Item Bank.

> **Nancy W. Burke** coordinated the loading of the item bank with test items by designing content specifications for categorizing items, writing item specifications, developing procedures for selecting or writing items and establishing quality control standards for item entry.
>
> **B. Darwin Kaufman** designed the item bank, wrote the specifications for the item bank system, conducted the process to acquire both the hardware and software, and coordinated all the actual development activities of the bank.
>
> **NOrman Webb**, as administrator of the competency-based testing program, led the developmental effort by providing both administrative direction and coordination including budget and policy development, recruitment and personnel assignment, interface with various agency functions in curriculum and instruction and technical assistance to local district users.

While these individuals assumed leadership roles in developing the Wisconsin Item Bank, many other members of the Department's staff in competency-based testing, pupil assessment, curriculum and instruction and data processing contributed significantly to the development of the Wisconsin Item Bank. Item Bank content development benefited from the generous advice of faculty at the University of Wisconsin School of Education. Finally, the support from Herbert J. Grover, State Superintendent of Public Instruction, was most important to the development of the Wisconsin Item Bank.

Description of The Wisconsin Item Bank System

Overview

The term, *item bank*, has been used to describe a variety of forms of test item collections. Three variables may be used to classify existing or developing item banks. These are:

1. the organization and structure of item collections;
2. the extent to which procedures for accessing, reviewing and retrieving items is systematized; and
3. the extend of the automation of procedures.

Using these variables, every item bank can be classified in a three-dimensional space.

The Wisconsin Item Bank is a system which uses a computer to store, review, and retrieve test items for the purpose of constructing tests. To understand this system, it is helpful to note how it varies from other item banks. This can be done by considering its classification within the three dimensions above.

Figure 1 below depicts the location of the Wisconsin Item Bank in a space formed by these three dimensions. On each dimension the Wisconsin system is located in the middle of the extremes. In the dimension of item structure and organization, the Wisconsin item collection has been reviewed carefully to ensure conformity to prescribed item writing standards. Currently, most of the items have not been field tested so the collection has not been classified by using any statistical information. In the dimension of systemization, the Wisconsin Item Bank requires users to follow a specified sequence of steps, but they have the freedom to make choices and special requests. In the dimension of automation, all of the functions of item retrieval and test construction are performed on a computer, but data management and statistical analysis are not automated.

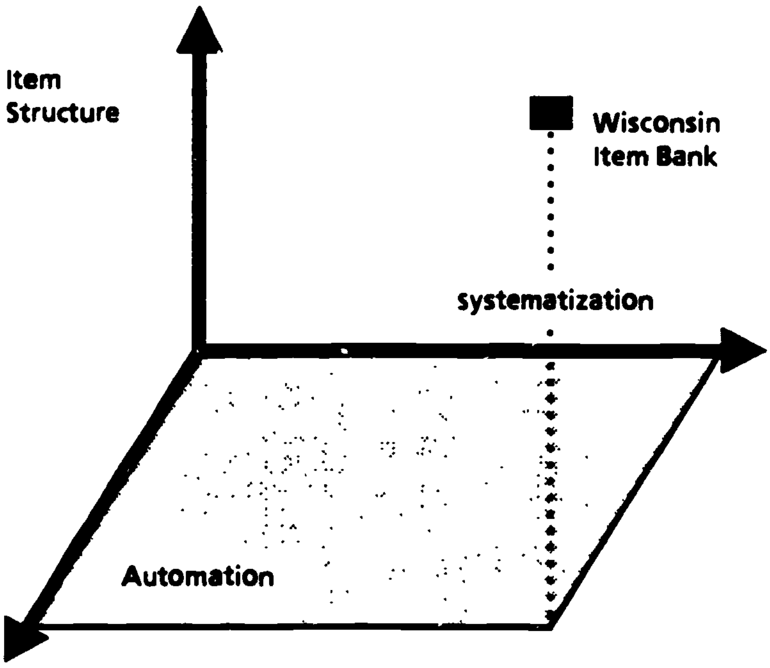


Figure 1: Item Bank Space

Purpose

The Wisconsin Item Bank was designed to accommodate a variety of district curricular definitions. The Wisconsin Legislature intended the Item Bank to provide local school districts with the ability to develop competency tests while retaining considerable latitude in the definition of the content to be tested. A hallmark of the Wisconsin Competency-Based Testing Program is that districts determine their own values with respect to curriculum and educational practices.

In order to use the Item Bank, district personnel must come to the Wisconsin Department of Public Instruction office in Madison to review and select test items. Prior to this, user districts identify the specific categories of the item collections which match their curricular needs. The activity during the visit is simply to review and select test items at a computer terminal. The selected items are then printed for further review at the district level. Finally, the district returns the exact specification of test content, and a camera-ready copy of the test document is printed and returned to the district. The process, including the local personnel involvement in the final selection process, takes from three to six weeks.

Computer System

The original design of the Wisconsin Item Bank called for a total computer system performing complementary functions. The envisioned system was to integrate the strengths of a text/image machine with those of a mainframe computer. The text/image capability was deemed essential to accommodate graphics, multiple type fonts, custom document production and high quality printing. However, existing equipment could not combine the text/image capabilities with other functions such as statistical analysis and data management. Therefore, a two-machine system was adopted. The success of this system design depended on the development of a linking facility which would unite the text/image capacity of one type of computer with the data storage and calculation capacity of a mainframe computer.

Even though this overall design was adopted by the Department in early 1982, it has yet to be fully implemented. At present, the Item Bank functions as a one-machine system with exceptional text/image capabilities. Since no link to the mainframe computer has been completed, all data management and statistical analysis is handled manually.

The existing system used by the Wisconsin Item Bank is quite effective for item perusal, selection and document production. Teachers and administrators can examine test items with attractive, effective graphics at a video display terminal, select those deemed appropriate and receive a test copy of publisher quality. From the perspective of the user, the capabilities of the Bank are excellent. The system's functional shortcomings, which impact on efficiency and item quality control, are transparent to the users. (See Figure 2 on page II-4.)

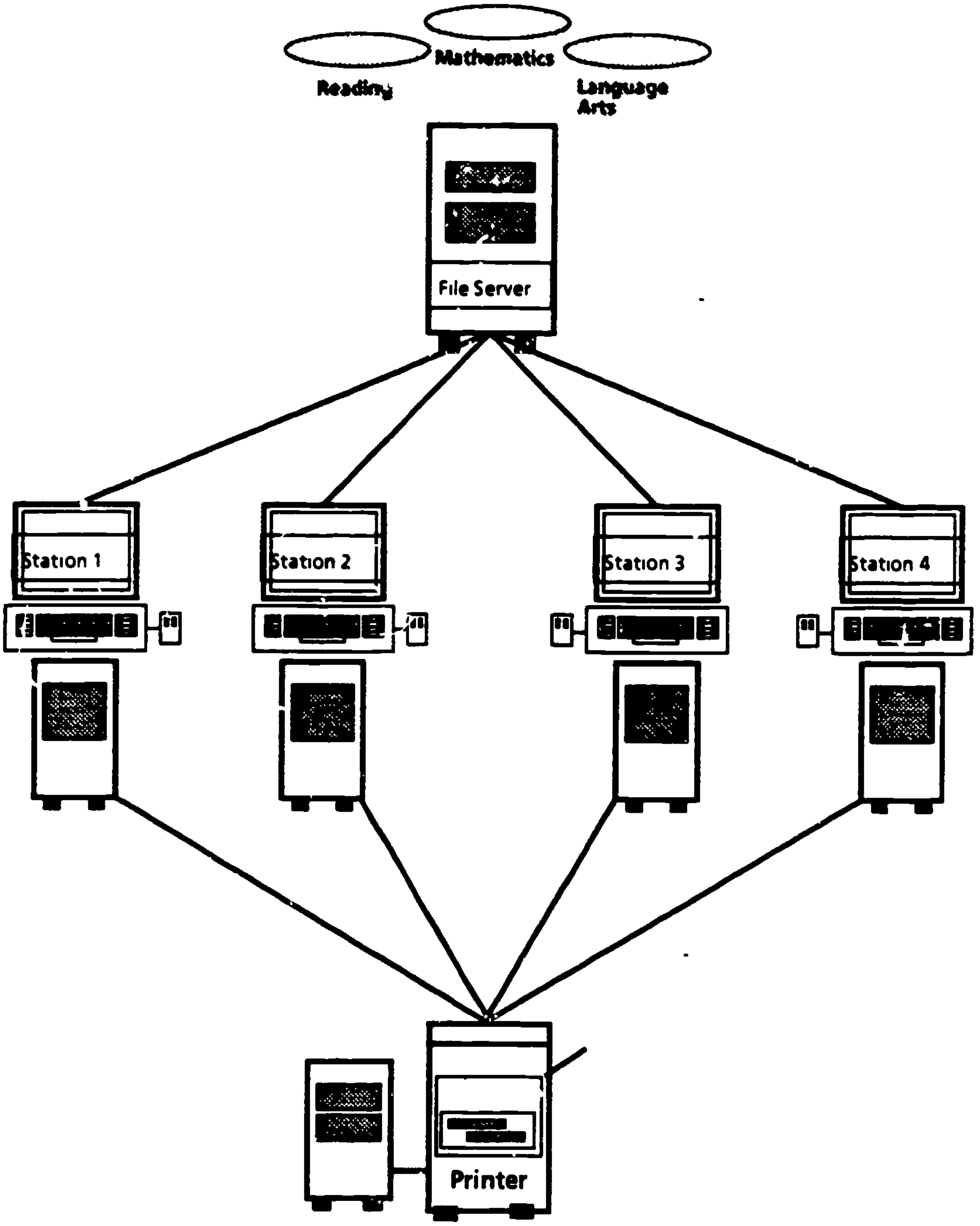


Figure 2: Text/Image Equipment

Three item banks -- mathematics, reading and language arts -- are stored and retrieved in print-ready form from a 300 megabyte file server on a Xerox Star System Network. Graphics and text are combined in both the display and printing of test items The Network includes four 42 megabyte Star Workstations and a laser printer.

Matching User Needs to the Item Bank

The procedures to use the Item Bank are straightforward.

1. Competencies must be written and matched to the bank's content classifications schema.
2. Items are perused and selected at a video display terminal.
3. The intial set of selected items is reviewed by curriculum experts at a district site.
4. Final test specificatons are determined.
5. Camera copy of a test document is composed and printed.

The access paradigm for most item banks is based on a one-to-one mapping of domain specification statements (usually objectives) to items. This approach often requires that users write (or re-write) curriculum in the language and formats of the item bank. In cases where an item bank's content specifications are statements such as objectives or competencies, the translation is often awkward and imprecise. These translation processes generally result in a distortion of content and, to some extent, unintended content appears on a test.

The access procedure for the Wisconsin Item Bank accepts the fact that there is no single *best* form for content specifications. Furthermore, there is not a one-to-one match between competency statements and test items which will be agreeable to all users. There is imprecision in the match between curricular domain statements and test items. The typical language used by curriculum developer/competency definers is ambiguous and does not delimit a unique set of test items. Therefore, in order for a test to reflect accurately the intent of content statements, curriculum decision makers must have latitude in the items considered for selection. Thus, they control the resolution of item/content match.

In the Wisconsin Item Bank, content specifications serve as an entry to the items in the system. However, the actual item selection is assigned to the district curriculum decison makers. This paradigm for access to the item bank preserves local control of test content and improves the likelihood of test content validity. While this approach to item selection and review requires a substantial commitment of time and human resources, the outcome is a more accurate match between content and the test. (See Figure 3 on page II - 6.)

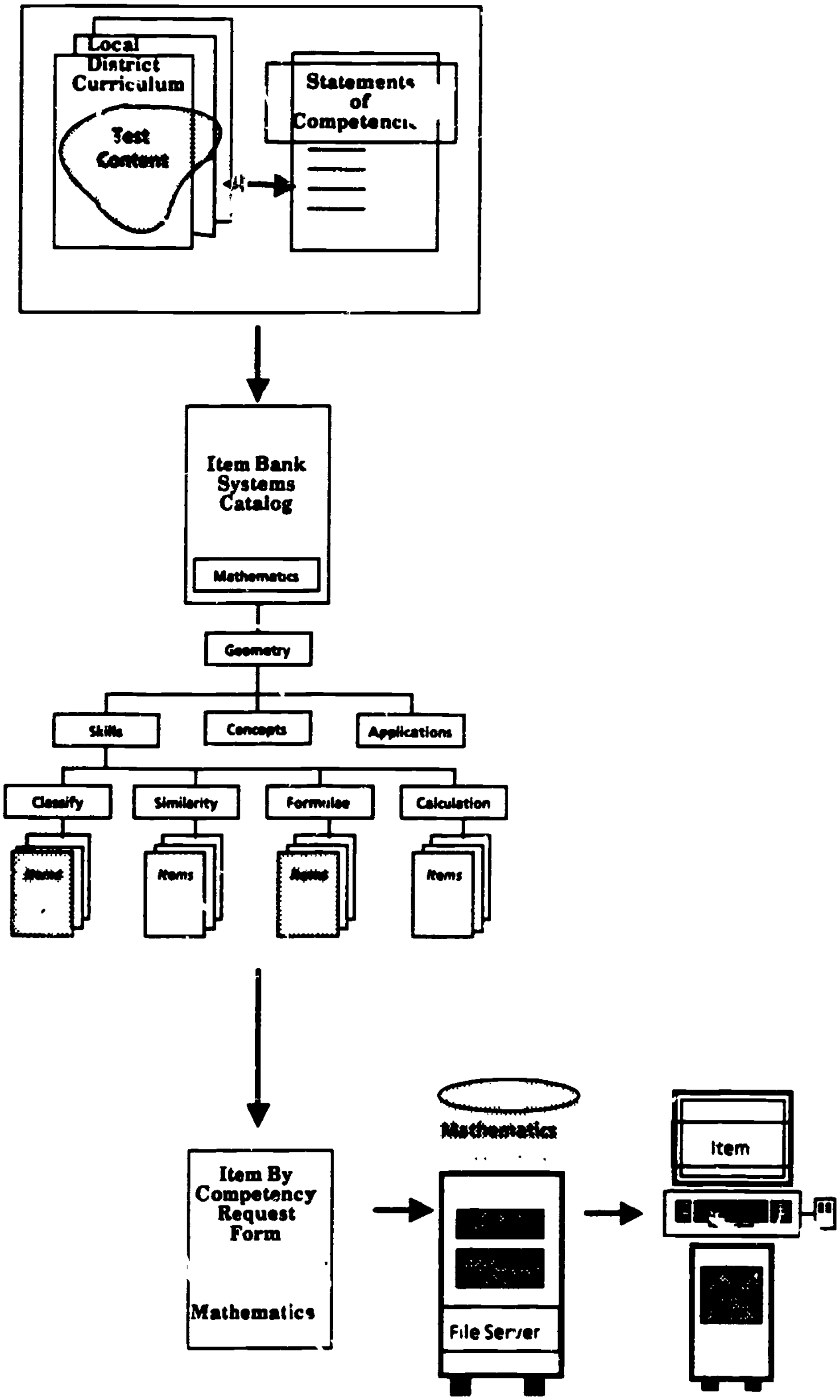


Figure 3: Matching Test Needs to Item Bank

Item Selection and Review

The process of selecting items for a test has two stages. During the first, items are perused on a video display at the Department's Madison office and a set of items which fit the curriculum are selected. These item selections are then printed for review by teachers in the district. The final set of items for a test are then specified and returned to the Department where equipment operators compose and print the final camera-ready test copy.

The most important criteria for choosing an item and, ultimately, including it on a test seem to be:

1. item/curriculum match
2. question format, and
3. language and content match with district instructional materials and style.

Clearly, decisions about each of these factors are idiosyncratic to each district involved in the process. The system's flexibility, which accommodates this degree of distinction, is a critical element in the success and utility of the Wisconsin Item Bank.

Test Composition and Printing

Once a district determines the final set of items for a test, word processing operators at the Department create a test document. This work is done on the same equipment used for reviewing and selecting the items. The operators' task is simply to call up an electronic copy of items originally chosen by a district, identify those finally selected for the test and electronically transfer those items to a test form.

The system includes a laser printer which produces publisher quality copy. From this single test copy provided by the system, each district arranges for multiple copy printing. Even though this is accomplished by a variety of methods, the results are nearly always of high quality. With an attractive camera-ready copy, and [illegible]n, relatively inexpensive copy processes, even the smallest district can produce first-rate test c[illegible]e Figure 4 on page II - 8.)

Summary

The Wisconsin Item Bank is systemized and extensively automated. At the same time, it is designed for maximum flexibility. A major design criterion, at all times, has been to encourage and facilitate district level decision making. This is a strength of the system.

Failure to implement the linkage for the original two-machine model has definitely been detrimental to future development. Without mainframe computer support, it is not possible to maintain adequate information about item performance for the 10,000 items in the Bank. Therefore, the capacity to improve item quality is severely limited. In addition, because there is no data management capability on the current one-machine system, the potential for growth is limited. Both the number of users who can be accommodated and the the number of items which can be included are restricted by the limiting characteristics of the one-machine test/image system.

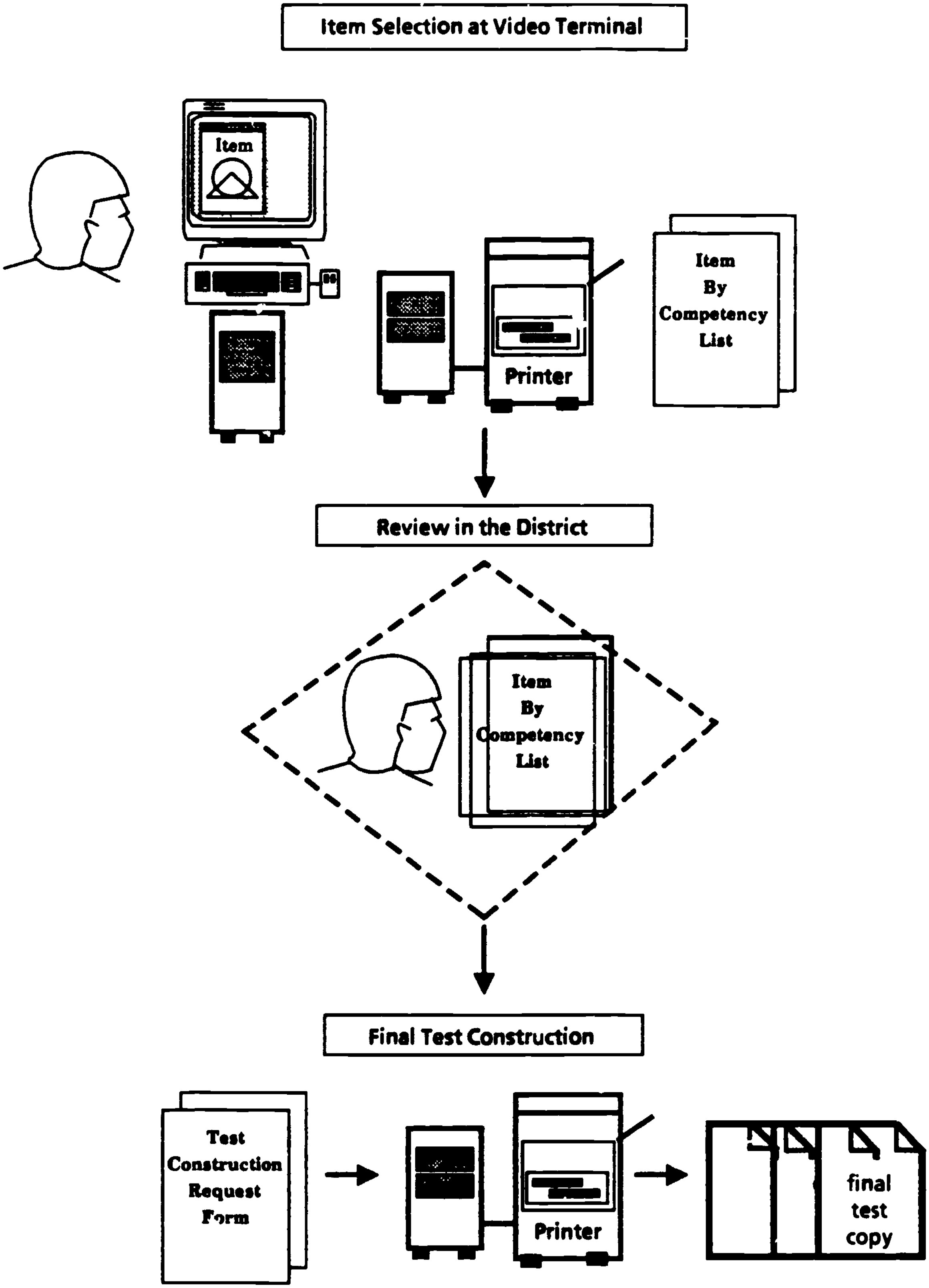


Figure 4: Item Selection and Test Development

Development of The Wisconsin Item Bank

Organizational History

The development of the Wisconsin Item Bank was, by no means, the result of a textbook design process. It began as a small, ancillary project within the Pupil Assessment Program at the Department of Public Instruction. Initially, its only proponents were the Program's director and staff members. In 1978, two item collections were obtained from the Northwest Evaluation Association and made available to school districts in Wisconsin. There were few users. During the Spring of 1980, a proposal was written and funded to develop computer support for the collections. An actual item banking system seemed to be on the horizon. Unfortunately, conflicts between two state agencies stalled the hiring of personnel and the grant money had to be returned to its source.

In 1980, two Wisconsin legislators became interested in competency-based testing. An aide to one of the lawmakers asked the Pupil Assessment Program staff about the requirements for a competency-based testing program. This legislative interest appeared to come from the perception that there were significant numbers of Wisconsin high school graduates who were not meeting employers' expectations upon entry into the work force. There also seemed to be a related belief that many Wisconsin children were progressing through the elementary and secondary school system without acquiring basic skills in reading, mathematics and writing. The legislative solution to these perceived problems was straightforward: to test children at junctures in their schooling, to identify those who fell behind, and then to require remediation.

From a Wisconsin educator's point of view, the proposed solution had drawbacks. For example, Wisconsin does not mandate curriculum from the state level so any one test would not equally match every district's curriculum. Therefore, if districts were to use a test to identify students who needed help, every district's test would have to be tailored to the local curriculum. These issues and many other testing problems were discussed by legislative aides and the Pupil Assessment Program staff during 1980. There was a sincere desire to draft meaningful legislation and to include information from the educational assessment specialists in the legislative process.

Developing a test item bank was not the primary objective of the lawmakers trying to advance competency testing legislation. These legislators viewed the development of an item bank as a minor part of the legislative package but did recognize it would help gain political support for competency testing because districts could develop local tests. There was a general concern among districts and others that information from a statewide competency-based testing program could be used for district-by-district comparisons.

The development of a test item bank was considered much more important by the Pupil Assessment Program staff who had been interested in the concept since 1978. The Pupil Assessment Program wanted to provide local school districts with a service that would allow creation of tests tailored to local curricular needs. While competency testing, as prescribed by the legislature, had some pitfalls, the lawmakers support did provide an opportunity to develop a test item bank to allow customized testing.

The first attempt to initiate passage of a *mandatory* competency testing program was begun during the 1981-83 biennial budget process. A state senator introduced the legislation through an amendment to the budget bill. The main part of the competency-based testing amendment to the 1981-83 biennial budget bill was withdrawn because of lack of support and arguments that policies should not be included in the budget bill. However, funding in the amount $437,600 for the development of an item bank was retained. The item bank, which had been viewed as a component of competency-based testing legislation, was enacted into law in the summer of 1981.

The Department was mandated "by July 1, 1983, (to) develop a computerized bank of test items which may be used to evaluate pupil competency in minimum reading, writing and mathematics skills."

Nine months after the passage of the law enabling the development of the Item Bank, a bill was passed establishing a *voluntary* competency-based testing program in Wisconsin. A significant feature of both of these laws was the preservation of the strong Wisconsin tradition of local school district control in curricular matters. The Wisconsin Item Bank was included in the competency-based testing law as one of four sources from which local districts could develop or adopt criterion-referenced tests. The competency-based testing law moved the operational date for completing the Item Bank to July 1, 1984. Funds had already been authorized in the 1981-83 biennial budget to purchase equipment and begin the development of the Wisconsin Item Bank.

Then the legislative process required the Department to develop fiscal estimates for operating the competency testing program. The Department had to estimate the number of test items to be developed and entered into the Bank. With only two years to get both the system and the content for a computerized item bank developed, the Department's decision makers established a base of 2,700 test items for the Bank. Even through many people involved in the funding process recognized that as many as 50,000 items would be needed to support a customized testing service for competency-based testing, this base figure of 2,700 items was used in developing various assumptions about the necessary funding levels for item bank development. There was also some caution on the part of Department decision makers not to promise extensive development until the system design proved workable and until volunteer school districts joined the program.

The original assumptions for the competency-based testing program served as the basis for various future disagreements about exactly how much funding was required to develop the Wisconsin Item Bank. In future biennial budgeting, the Department had difficultv justifying requests for increased funding when it could be argued by legislators that the Item Bank, which became operational on July 1, 1984 with nearly 10,000 items, shouid be sufficient.

Equipment Acquisiton Process

While institutionalization of the item banking concept was incremental and fragmented, the design process was quite logical and efficient Design alternatives were explored, hardware and software examined, and recommendations tendered and accepted. This process took place during a three month period and was carried out by Department employees on a study team.

Initially, the system requirements of the bank were defined. This was accomplished by identifying a relevant set of issues, analyzing each issue and generating the requirements from the conclusions of the issue analysis. It was decided that the Bank was to perform three functions:

1. test development,
2. test production, and
3. item maintenance.

The requirements prescribed the nature of the Bank's approach to accomplishing these functions. One requirement was that the district representative who interacts with the system must be a content expert in the domain in which the test is being developed. A second was that the system must allow for entering both item text and illustrations.

These requirements served as the basis for identifying and analyzing alternative design approaches. Two fundamental means of providing service were considered. The first was to contract with a commercial vendor who would provide a fully developed item banking capability to districts within the state. The second was for the Department to develop its own system, either through the acquisition of hardware or by utilizing existing hardware. In either case, software would need to be developed or acquired.

One task of the design alternatives study team was to examine existing item bank systems. Initially, ten systems were identified of which seven appeared to have promise. Representatives of the agencies and companies responsible for these systems were invited to present their products to the study team in the fall of 1981.

The most difficult task for the design alternative study team was to translate system requirements of the envisioned item bank into computer hardware and software terms. This was accomplished by carefully defining the specific functions to be performed and then specifying the hardware/software requirements necessary to accomplish each of the functions.

In analyzing the equipment requirements, five categories or types of systems were studied:

1. agency mainframe (and upgrades),
2. word processing systems,
3. micro-computer systems,
4. text/image processing, and
5. mini-computer systems.

A number of alternative design solutions were considered: three near-term or interim solutions and three long-term solutions. The study team agreed that none of the available systems should be recommended. There, in fact, were twelve functional requirements which none of the vendors could meet.

The study team then recommended a two-machine system utilizing a text/image processing machine and the Department's mainframe computer. In the written justification for the design concept the study team noted:

> The "DPI two-machine system," not only meets DPI-CBT requirements, it also gives the most flexibility to accommodate the as-yet-undetermined detailed system specifications, as well as the unpredictable outcomes of the legislative process..."

In early January of 1982, the recommendation for the two-machine system was adopted by the Department's budget and policy leadership. In retrospect, the two-machine system, allowing flexibility in both development and acquisition, may have provided for the survival for the item banking project in the unpredictable legislative fiscal process. The procurement process for the one-machine text/image hardware and software began immediately.

A request for proposal was written and issued by the Department in the spring of 1982. Xerox Corporation was awarded the bid to provide a system of workstations, software, laser printer, file server and the linking network. Two workstations, a laser printer and an 80 megabyte file server were installed in November 1982. This sytem was to support the storage of 50,000 items. After a year of system development, it was determined that the 80 megabyte file server provided insufficient storage capacity and a 300 megabyte server was added to the system. The Department also purchased two more workstations in 1984 so that three units would be available for local district users. One station was used for system development. The costs for purchasing the equipment, annual maintenance and supplies are provided in Table 1 on page II-12.

Table 1

Cost of The Wisconsin Item Bank Equipment at the Time of Installation

1 Print Server		
Hardware	$ 31,972	
Software	$ 1,540	
Total		$ 33,512
1 300 mb File Server		
Hardware	$ 54,145	
Disk packs (4)	$ 5,132	
Software	$ 3,420	
Total		$ 62,697
1 80 mb File Server		
Hardware	$ 37,195	
Disk packs (4)	$ 1,996	
Software	$ 8,289	
Total		$ 47,480
4 Star Workstations		
1 Unit Hardware	$ 12,369	
1 Unit Software	$ 5,733	
Total		$ 72,408
Network Cost	$ 2,180	
Total		$ 2,180
Total Equipment Costs		$218,277
Total Maintenance Costs (annual)		$ 31,960
Total Supply Costs (annual)		$ 2,000

Item Acquisiton Process

Acquisition of 'tems for an item bank was begun when the Department's Pupil Assessment Program purchased test items from the Northwest Evaluation Association in 1978.

Beginning in 1981-82, the Department expanded this effort by contacting state departments of education, local school districts and others known to have test item collections. Some of these educational agencies shared items from their collections in mathematics, reading and language arts. The final item collection totalled approximately 90,000 test items from 26 educational agencies and local school districts around the country.

This acquisition process included various arrangements for use of items from each source. Most sources printed item copies for a service cost and made no mention of use constraints. The collections came in every conceivable printed form and were organized by as many classification systems as there were sources. Few of the collections provided use statistics.

Item Collection Development

Content requirements for the Item Bank were subdivided by legislation into the areas of reading, mathematics, and language arts. A plan to catalog test items in each of these subject areas was developed with the help of curriculum specialists from around the state. These experts served on committees to develop a framework for indexing items which reflected the full range of the mathematics, reading and language arts curricula in grades 3 through 12. The result of these committees' efforts was the development of tree diagrams representing major topical subsections of each discipline. These subsections were further divided into skill areas which seemed the most logical in terms of curriculum scope and sequence.

These schema diagrams were translated into storage areas within the electronic folder structure of the text/image computer equipment. Each folder was coded in a numerical sequence which linked the item storage locations to the tree diagram cells designed for each curricular area.

All test items in the acquired collections were numbered and then randomly selected for evaluation to determine the amount of editorial revision required to bring each item into conformity with the Item Bank's structure. A subsequent pilot study showed that all the items in the collection would have to be revised to some extent to fit the requirements of the Item Bank. Extensive revisions were required in the language arts and reading test item collections. Even the items from the item collections which fit into the Item Bank classification scheme had to be edited and revised to provide consistent syntax and grammar. Many items were also out-dated and had inaccurate answers or illogical foils.

Item .vriters were hired to edit and revise items to fit into the Item Bank cataloging system and to describe those items in terms of item specifications for each storage category. In some areas there were no items among the collections to provide even a sample for developing item specifications. Consequently, new :pecifications were written. ^espite efforts at arranging and sorting items , some categories in the Item Bank remainded empty While item editing and writing activities were supported by education consultants in the Department, each item writer was also a specialist in a subject area and brought extensive teaching experi.ence to the item development process.

Items writers were trained to use the common writing conventions and editorial standards for multiple-choice test item design. Furthermore, item writers and editors were required to attend training sessions on developing sensitivity to bias in terms of gender and race. Rules for item writing included specific requirements for inclusion of men and women in various life-style roles and inclusion of minority group members in positive ways. Items focusing on extreme violence or religious beliefs were excluded from con :ideration.

The items were entered eler ›nically into the Item Bank text/image computer, and the item record form was printed for editinr .nd proofreading . Despite these initial efforts, many errors were found in the items. Items were reprinted and further evaluated after the first year of use by local school districts. Some items were corrected, foils were changed or graphics were clarified. In addition, information from Item Bank users provided insight into areas of the cataloging system which were illogical or difficult to understand. During .he second year, items were also developed for areas where no items had previously been entered. Designing of new items is continuing because users are requesting items which measure higher order thinking skills, such as mathematics problem solving and essay writing.

System Development

Development of the Wisconsin Item Bank system was guided by the original design specifications which defined five main functions:

1. item entry,
2. item selection,
3. test construction,
4. record keeping, and
5. scoring /reporting.

Putting the one-machine system into operation took 18 months. This development process required the coordination of content needs and the capabilities of the equipment to reproduce and store the items. Small scale system studies were conducted in each content area of the Bank. Mathematics was used as the basis for the first equipment prototype to evaluate procedures for users to interface with the Item Bank's computer system. This prototype included a few graphics items.

Pilot studies showed that the storage capacity of the original file server was insufficient and a larger file server was added to the system. New releases of software and hardware increased the user workstations' speed and local storage capacity. Additional workstations were purchased for local district representatives to use in selecting items. The one-machine text/image computer system provided storage for at least 25,000 items and expansion capability through the use of hard disk packs.

Pilot studies also showed that the item format would have to be more efficient so that users could view several items on one screen. Item records were shortened by deleting all unessential information and simply displaying each item with an identification number, grade level indication and difficulty value. This shortened [illegible] allowed the viewer to see six to eight items at a time and scroll backwards and forward [illegible] compare many items within a section of the bank.

The selection procedures for users were refined by discarding any written instruction telling them how to operate the equipment. For each district user, a member of the Item Bank staff explains and demonstrates the basic *"mouse-selection"* process on the equipment. After less than 10 minutes of explanation, users are able to view items from the Bank independently. This meets a major system requirement of user friendliness. Beginning in July 1984, this system provided for item entry, item selection and test construction for the Wisconsin Item Bank.

The present Item Bank system is not fully automated. Items are entered by data entry operators in a format which visually approximates the items as they appear on a test. Items can be transferred in a word processing environment to various user documents. The user can view the items on the screen but must manually record the identification numbers of selected items. Tests are constructed by electronically transferring items, but these moves have to be initiated by a machine operator. There are no programs whi[illegible] automate the movement of items from the Item Bank to user documents.

Requirements and specifications have been written for a record keeping system, but this function has not been implemented because it is beyond the capability of the text/image computer's hardware and software. Some effort has been made to link the text/image environment to the Department's mainframe computer, but the two systems are not readily compatible. The test items, including graphics, cannot be stored in the mainframe and then returned intact to the text/image computer. Any development of a data base management system for the Item Bank will have to involve another system.

The scoring/reporting functions are currently being provided by a contracted vendor using another system.

Personnel Time and Costs

The development of the Wisconsin Item Bank and its operation required the involvement of a number of staff members in the Department. Some of these personnel needs were met by staff from within the Wisconsin Civil Service System. Other personnel were temporary or limited-term employes (LTE's).

The quality of Item Bank content development was affected directly by the Bank's proximity to the University of Wisconsin which served as a source for highly qualified subject matter specialists who were available to work on a temporary basis. Table 2 on page II-16 lists the position types and the numbers of employes who worked on the Item Bank during the fiscal year from July I to June 30.

Table 2

Staff Resources in Full-time Equivalent Employes

Administration

Position	1981	1982	1983	1984	1985
Bureau Chief (Assessment)	.25	.10	.10	1.0	.10
Section Chief (Competency Testing)	.75	.33	.33	.33	.33
Section Chief (Data Processing)	.25	.25	.25	0	0

System Design and Coordination

Position	1981	1982	1983	1984	1985
Assessment Specialist	1.0	1.0	1.0	1.0	1.0
Systems Analyst (Data Processing)	.75	.75	.75	.5	0

Content Development

Position	1981	1982	1983	1984	1985
Assessme 1t Specialist	0	.5	1.0	1.0	.5
Program Assistant	0	.5	.5	1.0	.5
Word Pro :essing Operator (item entry)	0	.5	3.0	2.0	1.0
Assessment Specialist (Item writers/editors: LTE's)	0	.33	3.0	1.75	0

User Services

Position	1981	1982	1983	1984	1985
Curriculum Specialist	0	0	2.4	2.4	2.4
Assessment Specialist (service coordination)	0	0	0	1.0	.5
Program Assistant	0	0	0	.5	.5
Word Processing Operator (test construction)	0	0	0	1.0	.5

Record Keeping

The volume of items in The Wisconsin Item Bank, the number of possible tests that can be created, and the steps involved to select items and to construct tests make record keeping a necessity. Ideally, and with the current technology, much of the record keeping should be done automatically with the appropriate data base management software for the hardware being used. However, the Wisconsin Item Bank does not have such a data base management capability because it is using only the test/image computer, not the full two-machine system called for in the original design specifications.

Having adequate records to support the items in the Bank was a critical part of the design of the Wisconsin Item Bank's system. Record keeping was one of the five major functions identified for the system. During the first year of development after the equipment was installed, detailed specifications for records were written as the item formats were being developed. During 1984-1985, the first year of operation, some work was done to link the text/image computer with the Department's mainframe computer. Item numbers were transferred between the two systems and some sorting done on the mainframe, but resources were not available to develop this link any further.

In acquiring the equipment for the Item Bank, a high priority was given to finding a user friendly system which could handle both text and graphics. The desire was for the equipment to be both easy to operate by teachers to select items and sophisticated enough to enter, store, and retrieve graphics along with written material. While keeping records was identified as being important, it was less important than other capabilities. A system that could provide all functions was not available.

The objective was to get an operating system in place in less than two years that could be used by districts to construct tests. Therefore, development efforts were spent loading the Bank with items and developing a system for selecting items and constructing tests. Development of record keeping was delayed. It also seemed reasonable that after the Bank became operational more funds would be made available to upgrade the system and provide for some form of record keeping.

The original Item Bank specifications described two main types of records: item use records and item performance records. The item use records were designed to provide information about the frequency of item selection, identity of users, frequency of item use in final test construction and other types of item bank management data.

Item performance records specified data on each item including source, bias review information, entry date, grade level, correct answer, and notice of revision and editing. Statistical information on item performance was to include the percent correct for an identified grade level and/or a calibration index.

A new software release for the existing text/image computer is expected in early 1986. It may provide for some record keeping functions. There appears to be no funding to support developing the linkage between the two-machines as described in the original design specifications. The hope is that this new release will make the test/image computer the only machine necessary to operate the Wisconsin Item Bank.

User Services

Testing and curriculum specialists from the Department are available to school districts. Workshops for school districts participating in the competency-based testing program begin each fall and continue throughout the school year.These workshops provide technical assistance in developing and understanding testing procedures. They also make available education consultants in mathematics, reading, and language arts who help districts organize curricular objectives and translate these educational outcomes of the curriculum into testing competency statements.

When the district has defined curricular goals and developed testing competency statements, it can ask for an appointment to select test items from the Item Bank. District personnel review the Item Bank storage records and identify storage cells in the Item Bank which contain items that match their testing competency statements. Each subject area and grade level request for Item Bank service is reviewed by Department staff to be sure a district reviews all potentially useful items. (See Figures 5 and 6.)

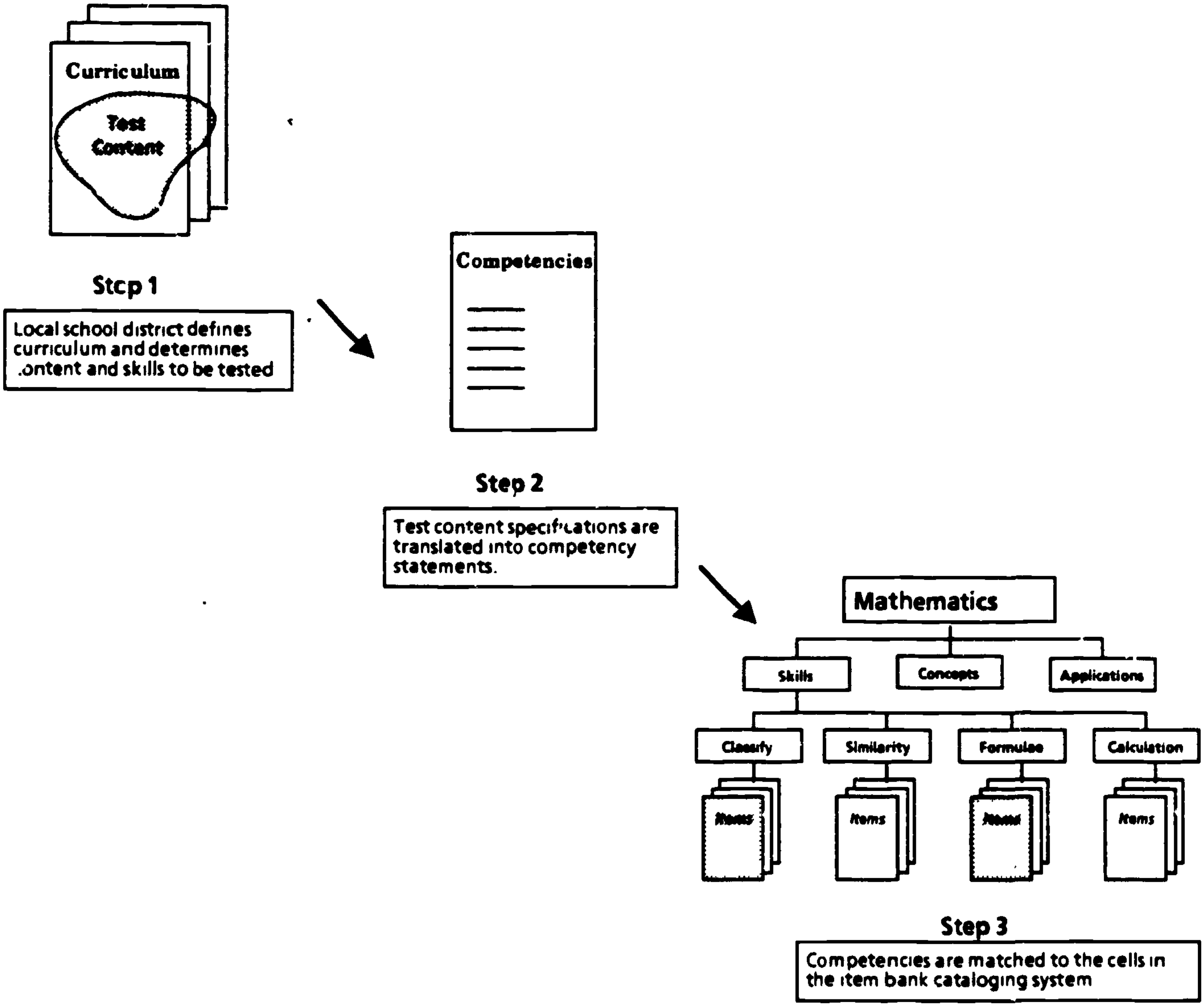
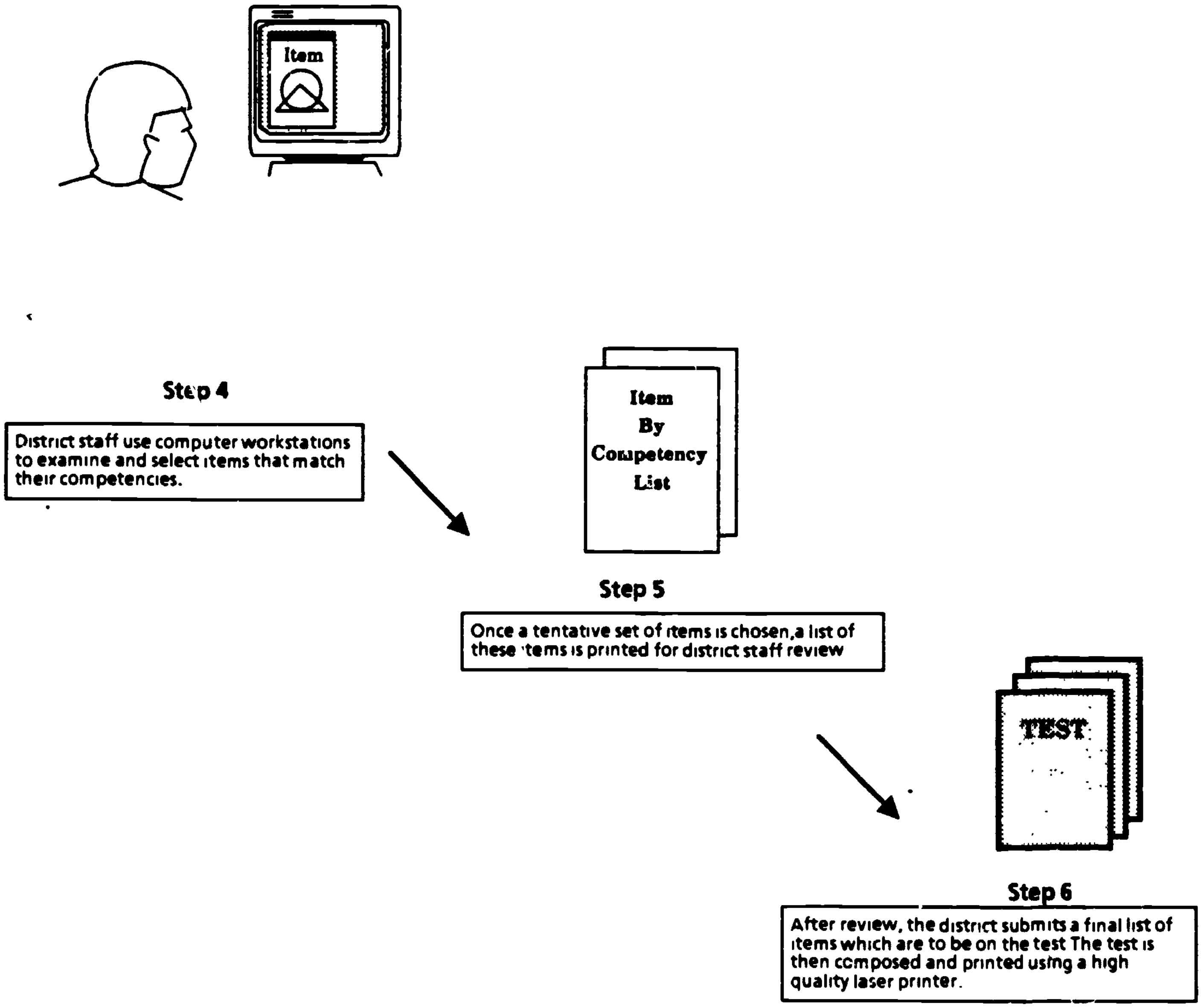

Curriculum
Test Content
Stcp 1
Local school district defines curriculum and determines .ontent and skills to be tested
Competencies
Step 2
Test content specifications are translated into competency statements.
Mathematics
Skills
Concepts
Applications
Classify
Similarity
Formulae
Calculation
Items
Items
Step 3
Competencies are matched to the cells in the item bank cataloging system

Figure 5: User Services

BEST COPY AVAILABLE

District personnel, teachers, and others review selected areas of the Item Bank and choose items which match each testing competency statement. This process gives the district users direct access to both the Item Bank and the support services available at the Department. Items selected from the Item Bank are printed displaying the items by test competency statement. This Item-By-Competency List is returned to the local district for final item selection. Using an electronic copy of the district Item-By-Competency List showing the items selected by the district, word processing operators at the Department format the final test copy.



Step 4

District staff use computer workstations to examine and select items that match their competencies.

Step 5

Once a tentative set of items is chosen,a list of these 'tems is printed for district staff review

Step 6

After review, the district submits a final list of items which are to be on the test The test is then composed and printed using a high quality laser printer.

Figure 6: User Services

Wisconsin Item Bank Problems and Issues

A review of the Wisconsin Item Bank development shows that many issues were resolved in getting the Bank operational. However, many problems still remain. Some of these problems may be unique to the location of the Item Bank in a state education agency but others are more generic. In conclusion, these problems and issues are presented for consideration.

Legislative and Organizational Commitment

The development of a functioning item bank requires a long term commitment of time and resources. Placing the Wisconsin Item Bank in a state governmental agency makes it vulnerable to changes.

The early stages of item bank development were carried out with no legislative interest and little organizational commitment. The impetus for organizational support came from legislative initiatives. At present, from the Legislature's perspective, the Item Bank is operational. The problems that remain are not the concern of lawmakers. If they are to be solved, resources must be generated by the Department. It remains to be seen if the refinement of the Bank into a quality service will be a priority within the Department. The answer to this will depend, in part, on its utility as perceived by districts.

Certainly, the Item Bank is institutionalized within the Department's bureaucracy. There is expensive equipment whose sole purpose is item banking. Several positions have been added to the Department to perform item banking functions. Also, there are a number of budget items for the support of the Bank's program. Furthermore, a number of school districts have come to rely on the Bank to support a portion of their testing programs. With equipment, personnel, budget, and user expectations, operation of the bank for the foreseeable future seems assured. However, there does not seem to be organizational or legislative support for growth or expansion. Funds and personnel for such efforts will have to come from imaginative management of existing program resources. A challenge to the Item Bank leadership will be to maintain the excitement and enthusiasm of its beginning as the program matures. The reward system of the bureaucracy will support the status quo. It will take unusual personnel to push the project to its potential.

Quality Control

Maintenance of a quality item bank requires that statistical information about item performance be gathered and interpreted on an ongoing basis. Content validity is simply not sufficient. Gathering such information can be done through formal field tests or as the items are used on local tests. The former is demanding, both in terms of dollars and the number of subjects required for the volume of items in the Wisconsin collections. Using data which are gathered from local testing is a feasible and inexpensive way to gather statistics for item quality control. To implement such an approach requires that test scoring be integrated into the Item Bank's system so that item records can be automatically updated as new data are gathered and analyzed.

At present there is no such integration. Further, the lack of a data management capability makes the logistics of the necessary updating impossible.

While the capability to aggregate information about items now could be accomplished with data which is available to the Department, updating item records must await the development of software for the text/image machine or the implementation of the two-machine system.

Vendor Dependence

The development of certain functions or expansion of the Item Bank may be finally limited by not having full access to the text/image computer's hardware and software.

The Wisconsin Item Bank operates on a "*closed system*". The Department, like other customers does not have access to any computer code and therefore cannot modify or write software for the equipment. This presents two significant limitations for the Bank's development. First, the Department must wait for the vendor to release needed programs. An example is the data management system. A second problem, which stems from the proprietory nature of all codes, is that data which have been entered on the existing system cannot be transferred to another company's machine. Because there are over 10,000 test items consuming nearly 300 megabytes stored by proprietory code, it is not feasible financially to switch to another brand of equipment.

At least a partial solution to this problem lies with the two-machine system. Since the vendor is establishing compatibility with mainframe computers, there is hope that packets of data can be managed on the Department's mainframe and returned to the text/image processor for interpretation. Such procedures would not require the mainframe computer to interpret data, only to identify and manage it. With such an approach, all management item data would take place on the mainframe computer, while user access would be at the text/image terminal. From a user's viewpoint, the location of processing would be transparent.

Graphics Entry

From the beginning of the development process, item graphics and illustrations have presented a challenge. Initial decisions about the nature of equipment needs were based on the belief that illustrations must be displayed along with text as users view items on a video terminal. This early decision has proven to be a wise one. However, it has been costly in terms of the effort required to enter graphics and illustrations, into the system.

At present graphics are drawn at a workstation using a "*mouse*." Even with excellent operators, the most simple drawings are tedious to reproduce. Though the initial specification for the system envisioned digitized entry of graphics, that technical capability has yet to be realized.

While entry of graphics into the system is essential, it is also expensive. When the time comes that pictures can be prepared by a graphic artist and then scanned into the system and merged with text, item entry will become much less expensive.

Information Management

Nearly all manipulation of items within the Wisconsin Item Bank's system requires operator intervention. For example, when items are chosen from the review set for inclusion on a final test, an operator must take action to move, electronically, each item on to a final test document. The text/image system does not have an effective means of merging a data base of items containing graphics. The implications of this are twofold. The system is costly to operate because of personnel requirements and has limits with respect to both number of users and items. Both the capacity and the efficiency of the system would be enhanced considerably by an effective data management capability which would handle graphics as well as text.

When such a management system is available and implemented, the Item Bank will become extremely efficient. By using prompts and menus, a user with 15 minutes of training would be able to perform all necessary operations to the point of actually composing the final test document. Even this final action would be greatly simplified by an effective data management system. Automated data management would easily reduce current operating expenses to one-fourth the current level.

Remote Access

A maior obstacle to use of the Wisconsin Item Bank is its location. The majority of school districts are more than 100 miles from the site. Considerable discussion and study has been devoted to this problem. Although there are issues of item security and maintenance, the overriding problem is cost of both equipment and personnel.

As presently operated, any terminal site would require at least one full-time person. This, added to the expense of equipment ($15,000 -$20,000 per site), makes such expansion prohibitive. However, in the future, with the addition of sophisticated data management capability and anticipated reduction in cost of equipment, remote access should be feasible. Personnel requirements should be reduced to a 1/8 or 1/4 time clerk and initial purchase cost of equipment should be well below $10,000.

As cost becomes tolerable, it is anticipated that each of the 12 intermediate education units in the state would have an item bank capability. Further, many districts will likely choose to acquire their own access to the Wisconsin Item Bank.

With the expansion of the item collections into other curricular areas, a proliferation of access into the file would make the bank a major resource for instruction assessment in all Wisconsin districts.

Customization of Test Items

Recent experience shows that some users of the Wisconsin Item Bank want to create customized test items as well as customized tests. Even with a choice of two or three item formats in the Bank, some districts ask to change test item wording to match a particular format unique to classroom instructional methods or district curricular practice. While the users acknowledge that some changes may pedantic, the need to express individual local curriculum in test item construction is strong.

The Wisconsin Item Bank has attempted to accommodate these special requests of users, but it takes considerable time to reformat items on each test and any statistical information which might be provided with an Item Bank item is not longer valid for district use. Users who are concerned about item performance data may decide to accept Item Bank formats while others prefer to field test customized items and retain them on test forms.

There are two types of problems that result from districts modifying items from the Wisconsin Item Bank. The first involves the issue of psychometric validity. Districts cannot use statistical information provided on the original item forms. The capacity of an item bank to provide item performance data is one of the primary reasons to use the service to create customized tests. Another problem with customizing items is that it interferes with the Bank's test construction process. The system is not efficient when an operator has to key enter customized items on to an Item Bank test form. The test form has been designed to accept item formats from the Bank.

If the Wisconsin Item Bank is going to provide customized item development, it will have to develop a system to manage it without exhausting staff resources which are committed to districts using items as they appear in the Bank. While the Item Bank service should be sensitive to the individual user's needs and flexible enough to accommodate local district customized items, it must develop procedures to insure the integrity of Item Bank use and communicate the limitations as well as advantages of developing customized tests with items from an item bank.

Conclusion

The Wisconsin Item Bank is a functioning system providing customized tests for school districts participating in a statewide competency testing program. The development of the Item Bank is continuing as problems and issues are resolved in an ongoing effort.

District users of the Item Bank have had an opportunity to evaluate what is important for students to master in their local curriculum. Teachers have acquired an increased awareness of the link between their curricular objectives and the content of tests. It is not possible to evaluate the benefits of customized testing for the students and parents because the first testing cycle is just beginning.

When the first customized testing cycle is completed, the Wisconsin Item Bank must be evaluated to determine its strengths and weaknesses. Ultimately, the Wisconsin Item Bank is a valuable service, if tests from its system accurately and efficiently identify students who are not mastering important aspects of a local district's curriculum, and if these tests provide direction for remediating deficiencies.

III.

A REGIONAL AND LOCAL ITEM RESPONSE THEORY BASED TEST ITEM BANK SYSTEM

Walter Hathaway, Ron Houser, and Gage Kingsbury

Portland Public Schools
Department of Research and Evaluation
Portland, Oregon

November 1985

INTRODUCTION

A decade and a half ago, a small group of researchers and test developers in the Portland, Oregon School District realized the potential of Item Response Theory for developing calibrated item banks that would be tools for better educational measurement. The purposes we wanted our citywide testing programs to support included equitable, effective, efficient, valid, and reliable:

- Grouping and placing students
- Targeting instruction on individual student learning needs
- Evaluating student progress over time
- Identifying neglected areas of the aligned curriculum and evaluating and improving programs and services at the student, classroom, grade within school, and grade within district levels
- Providing accountability to the school board and the community

In order to meet these needs, we needed an educational measurement system that would answer the following questions:

- Is the current rate of gain of this student, class, grade, or program satisfactory compared to his/her/its age, grade, program mates and the previous pattern of gains observed?

- What are the current strengths and weaknesses (in terms of goal areas needing further diagnoses and possible work) of this student, class, grade, school, or program, and how have they changed over time?

- Is the level at which this student, class, grade, or program currently performing satisfactory compared to his/her/its age, grade, or program mates and the previous pattern of levels observed?

We could not find any available measurement p ıgram that would help us answer these questions and meet these needs adequately, and so we set out to build one ourselves.

There followed a period of extensive collaborative research and development, much of it within the framework of the Northwest Evaluation Association, which was created to foster regional cooperation in and mutual benefit from this effort. The result today is a system of three comprehensive basic skills item banks in Reading, Mathematics, and Language Usage. The constantly growing item banks in Reading and Language Usage each have over 2,000 field tested, calibrated items linked to a common, continuous curriculum scale for each subject. The Mathematics item bank now has over 3,000 such items. State and local school systems have been using these item banks since 1977 to construct effective, efficient survey achievement tests, competency tests, and other instruments that combine the best qualities of criterion referenced and norm referenced measurement. These excellent measurement systems have been the cornerstones of state, district and school renewal efforts that anticipated "A Nation at Risk" (Gardner, 1983) by at least five years. The ongoing collaboration is now resulting in a similar item bank in Science, and yet another in Social Studies is on the drawing board. Computerized adaptive versions of the tests based upon the item banks are now being pilot tested in Portland.

DESCRIPTION OF THE CURRENT ITEM BANKS

At present, there are approximately 2,200 items that have met all the pre-screening, editing and statistical criteria in Reading; 2,300 items in Language Usage; and 3,500 in Mathematics. Each year 1,800 new items are field tested across the three subject areas, with about 1,200 being added to the item banking systems. Items are added to the banks by a process called the "Developmental Testing Program." Within two weeks of sitting for a regular Achievement Level Test, all students in grades three through eight also sit for a short test of between 20 and 25 items in either Reading, Mathematics, or Language Usage.

In an attempt to more reliably calibrate items on a short test with the current item banks, a procedure is employed that is termed the "fixed parameter model" (Houser, Hathaway & Ingebo, 1983). Since the students have a relatively reliable estimate of their ability through the Achievement Level Testing Program, their ability estimate can be essentially "fixed" for the Development Test at that point on the underlying metric. Item difficulty is the only parameter that is necessary to estimate. Thus, one can calibrate a one item test with the same precision as is achieved on a functional level test.

In cooperation with the Northwest Evaluation Association Science Project, Portland Public Schools is currently in the process of developing a Science item pool which now consists of about 3,500 items.* These items have not been scaled to a common metric at the present time; however, the project is currently engaged in the first phase of a research project that will help determine the number of scales needed and allow the project to begin the design of a series of field tests leading to a final set of scaled Science item banks. The project has plans to scale as many as 16,000 items by the year 1990.

WHY THE ITEM BANKS WERE DEVELOPED

Advantages of Item Response Theory

Classical test theory has been the backbone of educational and psychological testing for most of this century. The concepts of reliability and validity have enabled the development of tests of psychological traits based on sound theory and practical experience. In the last two decades though, testing techniques have been developed and refined to use new theoretical models to describe the interaction between the test-taker and the individual test questions.

*The term "pool" was used to distinguish the set of items which do not have a common metric. The term "bank" has been reserved for items which are scaled to a common metric.

These item response theory (IRT: Lord & Novick, 1968; Lord, 1980) models allow the estimation of differences in the measurement characteristics of test questions. These item characteristics may then be used to determine an individual's score on a test and the precision of the score.

The major advantage of these IRT models over classical test theory is that classical item and test characteristics vary, depending upon the group of students taking the test, while IRT item and test characteristics do not. Classical indices of item difficulty, point-biserial correlation, and test-retest reliability may all change when groups of test-takers differ in mean ability or ability distribution. When IRT models are used, item parameters are not biased by the specific sample of students taking the test or by the other items included on the test reference. In addition person parameters are not biased by the particular form of the test taken or by the other students taking the test.

This sample independence means that, except for differences in precision, the student samples used to estimate item characteristics are interchangeable, and the item samples used to estimate student characteristics are interchangeable. It should be noted, however, that these properties of IRT only pertain if the items and students are sampled in some manner from the populations of interest used to derive the IRT scale (Lord & Novick, 1968).

A second advantage of IRT over classical test theory is that a measure of the precision of measurement is available for any question or set of questions with known item characteristics. This standard error differs from that found in classical test theory in that it varies as a function of the student's achievement level. This is a much more rational result than the constant standard error obtained in classical test theory since we know that the error is not the same for individuals scoring, e.g., 10, 25, and 40 on a 50-item test.

The result of these properties of sample independence and a known functional form for the standard error is that test developers can develop test forms with confidence in the comparability of scores and with some knowledge of the precision of the scores that will be observed. Any set of items taken from a bank of items calibrated using this procedure will produce a scale score and correspondence error (measurement. The size of the error depends on which items are selected, how many items are selected, and the student's raw score on that set of items.

Item Banks and IRT: A Common Scale

A second consequence of the properties of the IRT models is that large groups of items addressing the same subject can be brought together onto a common scale. This makes the IRT models invaluable to anyone trying to create or maintain a large item bank. Large item collections may be easily brought together for use with classical test theory (as in domain-referenced testing); but in order to create anything but randomly parallel test forms, some type of overall scale is a necessity.

The linking procedures and designs which are a direct outgrowth of IRT theory allow researchers to scale thousands of items together (Hathaway, 1980).

From these item banks tests for different purposes and measurement of students' different functional levels may be drawn. Without this common scale the usefulness of a large collection of test questions is severely limited.

THE DEVELOPMENT AND EXPANSION OF THE ITEM BANKS

In the early 1970's, the Portland School District's Research and Evaluation Department, inspired by the work of Georg Rasch and Benjamin Wright, began investigating the applicability of IRT models to the development of item banks from which achievement tests in Reading, Mathematics, and Language Usage could be produced. The Rasch model or one-parameter IRT model was identified as the most promising of these models at the time because of its more advanced stage of development as well as its conceptual clarity and practicality. Later in the decade, while other test developers moved to two- and three-parameter models, Portland's Research and Evaluation Department concentrated its efforts on improving methods of linking items in Rasch-calibrated tests to a single, continuous scale; on reducing the need for second-parameter correction by discarding items that did not fit the model; and on reducing the need for third-parameter correction by readministering higher or lower level tests to students not performing at "percent right" levels that fell within the measuring range of the tests.

Before accepting the Rasch model as a useful tool for the construction of item banks and tests, the Portland Research and Evaluation Department conducted extensive research on such questions as:

1. Does the calibration procedure yield the same scale values regardless of the student sample?

2. How many students are needed to scale test items within acceptable error limits?

3. Does the calibration procedure used produce comparable achievement levels for the same students on different tests?

4. Does an item receive the same scaling regardless of other items in the test?

Once the initial basic research had been completed, work began in earnest on collecting, writing, screening, cataloguing, field testing, calibrating, statistically analyzing, and linking items together into true item banks. Impetus was given to planning the content coverage of the banks by an earlier regional project called the Tri-County Course Development Project which developed, classified, and catalogued comprehensive sets of K-12 learning outcome statements in 14 subject matter areas including those covered by the first three item banks. Collaboration by state and local school districts in Oregon and Washington was gained within the framework of the Northwest Evaluation Association which proved a rich source of ideas, items, field test sites, and other critical, major contributions.

Item Sources

There are two procedures that are employed to identify new items to be field tested. First, items are identified from existing sources such as the National

Assessment of Educational Progress released items, items developed by other federally supported projects, and items shared by other state and local school systems. These sources were an extremely valuable and relatively inexpensive way to get potential new items in the early stages of the item bank development. Prior to the selection of items for each Developmental Testing Program, goal areas which are either lacking depth or some ranges of difficulty are identified. Over time, these "weak" areas have become fewer and thus the specifications for new items have become more restrictive. As this has occurred the public domain item sources have not been as productive a resource as they were in the initial phases of bank development.

Second, items are written to the specifications by trained item writers who usually are teachers. These teachers normally have considerable experience in teaching the level of students for whom they are to write new items. During the item writing process, teachers have the opportunity to critique one another's items. This process helps refine the end product, particularly when writing to a higher level of thinking goals. Considerable care is taken to help the item writers focus on district course goals rather than specific behavioral objectives. It is a policy of the Research and Evaluation Department that specific teaching methodologies not be measured. Rather, the more comprehensive learning goals to which those methodologies are directed are the guiding organization behind the development of the district-wide measurement program. The particular teaching strategies that should be employed in the classroom to help students meet the system's educational goals are left to the professional judgment of the individual teacher.

A number of format specifications are applied to screened and newly written items to ensure good item writing practices and to meet the need to have consistency among the items in a given bank. The sorts of things these guidelines address include: easily understandable directions, clarity, item responses of relatively the same length, good distractors, only one correct answer. If these specifications are not met, the items are either returned to the item writers or item screeners or, if the problem is a minor one, is simply modified by the item bank technical staff.

Items which meet the screening specifications are then assigned an identification number and are sent to the typist. Once the items have been initially formatted on a word processor, they are checked by a diverse team of professionals to determine whether they have any disqualifying gender or ethnic problems identified from an extensive checklist. Problem items are discarded or revised by the team and returned for correction. A final edit is made after all other corrections for any remaining format, spelling, grammar, and/or punctuation errors. The items are then grouped with other items assigned to the same grade level by the item writers or screeners.

Groups of items are then formatted into developmental field tests of 20 to 25 items and administered to students. Once the students' item responses are received, they are matched by student identification numbers to their Achievement Level Test score for that subject. A series of checks are made to determine if the match is correct and if the students did indeed respond to the form of the test indicated on their answer sheets. Upon satisfactory completion of these tests, the "fixed parameter" calibration analysis is run.

Item records are developed for each item on the test containing: the item number on the test, the item bank number, the keyed response, the test form number, the number of omits, the number of students responding to the item, the test date, the Rasch calibration, the calibration standard error, the point biserial correlation, the mean square fit, and the percent of students answering the item correctly. Item characteristic curves are also developed and printed for later evaluation.

The first item screening criterion used following testing is based on the number of students responding to the item. If this number is less than 200, the item is flagged for retesting in the next Developmental Program. The next criteria concerns the percent of students correctly answering the item. If this is less than 5 percent above chance level or greater than 90 percent, the item is flagged for retesting at either a higher or lower grade level as appropriate. In either of these two extremes it is felt that not enough of the item characteristic curve (ICC) is available to reliably determine its calibration. Next, the mean square fit (MSF) is checked, and if there is not a statistically significant misfit, it is then passed to the final statistical check. The last check is a visual inspection of the ICC to identify any abnormal deviations from the ideal model that was not picked up by the MSF.

Any item which fails the first two tests is automatically scheduled for retesting. Any item which fails the last two tests is sent back to a team of item writers for revision if the problem can be identified. If no obvious problem is identified, the item is deleted.

HOW WE CURRENTLY USE THE ITEM BANKS

Test Development

The Reading, Mathematics, and Language Usage item banks are the prime resource employed in the development of the various functional levels tests administered to all students in the Portland Public Schools, grades three through eight, in both Fall and Spring (see Haney, 1985, for a review). Two series of tests in each subject area are designed to measure all six grade levels except in Mathematics, where the instructional sequence of the curriculum dictated the development of two series for each Fall and Spring program so that high achieving younger students would not be expected to respond to items measuring goals for which they have had no instruction.

Each series consists of multiple-level tests of different difficulty. Each level is 20 RITS (2 logits) wide in difficulty. (For students in a particular grade, the standard deviation of achievement levels is about 15 RIT points.) A level's content overlaps the previous and next higher level by 50 percent. A minimum of seven items is necessary to report a subscore and a total test consists of between 45 and 60 items. Students are placed into these levels by either a short locator test, teacher judgment, or by any or all of their previous six scores.

The item banks also comprise the primary resource for the construction of the Graduation Standards Testing Program (Hathaway, 1980). These tests are designed to make competency decisions and so are focused on a cutoff point on the underlying metric. Most of the items are centered around the cutoff point so that the best measurement is obtained at that point. However, an extension of the measurement downward is achieved by including items which are less difficult. This is done so that goal-based information is available to help identify weak areas for students not performing above the cutoff point.

Computerized Adaptive Testing

To allow more school-based control of testing and to provide school personnel with a capability to use the techniques of modern test theory, we have been investigating the uses of computerized adaptive testing (CAT: Weiss & Kingsbury, 1984).

CAT is a technique for testing which employs a computer to present the most appropriate test questions to a student. The student answers questions on the computer keyboard and the computer scores and records the student's responses. After each item the computer searches an entire item pool (previously stored in the computer) and selects the next test question for the student. At each step in the test, the computer refines its estimate of the student's achievement level and gives the student the question left in the item pool that will provide the maximum information.

From the student's point of view, the test questions become more difficult as questions are answered correctly and less difficult as questions are answered incorrectly. This is akin to the levels assignments done in the functional levels tests described above, but in a more direct manner, based on the student's current performance.

The CAT system will help serve the schools' special needs for interim measurement, initial pupil placement, and individual graduation standards testing. It is hoped that this form of testing will allow teachers and administrators to assess gains from individualized instruction that occur between the district-wide assessments that occur in the Fall and the Spring. It is expected that the CAT system will serve as a useful addition to the current testing program and help us come one step closer to a continuous mea:urement model.

The present pilot project, active in six schools ranging from elementary to high schools, is designed to assess the reactions of teachers and students to CAT, and to provide assurance that the CAT system measures the basic skills in the same manner as the current, district-wide, paper-and-pencil tests. Future research will address the potential for enhanced measurement with the CAT system.

PROBLEMS AND ISSUES IN ITEM BANK DEVELOPMENT, MANAGEMENT, AND USE

Choice of an IRT Model

Although the one-parameter model is currently being used with the basic skills item banks, we have not rejected the possibility of the use of other item response models in the future. As mentioned above, the one-parameter model was originally chosen for use because of its strong theoretical and practical development. If another model were found or developed to the point at which it was demonstrated to have substantial, practical advantage over the one-parameter model, it could be used to recalibrate the current item banks or more likely be applied to the item banks currently under development.

Aside from the one-parameter model based on item difficulty, researchers have developed a two-parameter model which adds an item parameter allowing differences in item discriminatory power and a three-parameter model which adds an item parameter allowing differences in the lowest expected probability of a correct response.

These models differ in three respects. These are the estimation procedures used for item calibration, the scoring procedures used for estimating student achievement levels, and the extent to which student test responses conform to the model.

When building an item bank, it is essential that the response model chosen have practical procedures for the estimation of item and student characteristics. At the same time, the model should be flexible enough to capture the differences inherent in students and items. The one-parameter model has the edge in simplicity of estimation, particularly for student achievement levels. However, the more complex models may result in a more precise representation of item and student performance. Each item bank developer must weigh the advantages of each model with respect to the particular testing situation to determine the best IRT model.

When the basic skills items were first developed, a primary consideration was the number of students that would be needed to calibrate a given set of test questions. It became readily apparent that the only IRT model that would allow stable calibration and linking of items, and not overly restrict the growth of the item pools, would be the one-parameter model. This model allows stable and consistent estimation of item parameters with samples of 200 students. The two- and three-parameter models have required much larger sample sizes to obtain stable calibrations.

Recently, though, Swaminathan and Gifford (1985) have developed new techniques for estimation in the two-parameter model which may reduce the number of students needed to adequately estimate item parameters. In addition Lord (1983) has theoretically demonstrated improved measurement precision using the two-parameter model for all sample sizes greater than 200.

These studies as well as other research findings have led us to begin developing research designed to compare the practical impact of model choice on basic skills measurement.

Dynamism

To insure that the item banks continue to provide measurement reflective of the curriculum as it is being taught at any given time, the banks must be periodically checked and updated to correspond to these changes. There is some concern about the comparability of scores across time with the use of such a dynamic scale. However, metric comparability can be achieved by employing those subsets of items which have calibrations that remain stable over time. The total measure would still be reflective of the underlying trait at any given point in time.

By adding new items to the item banks after every testing program and by including at least two testing programs in the annual calibration check, the entire item bank would slowly reflect major curriculum shifts. Calibration drifts over the years have generally been modest. The second district-wide administration of the Language Usage test did result in substantial calibration drift, however. To explain this finding, the hypothesis has been advanced that teachers were, for perhaps the first time, focusing instruction on the Language Usage goals.

Interaction with Curriculum

We were fortunate in Portland and in the Northwest in general to have a tradition of a commitment to goal-based instruction and measurement that extended back to the Sixties. We also had a history of strong curriculum and evaluation leadership and collaboration.

One of the greatest challenges, however, in developing the item banks and using them to create good, useful, curriculum aligned, instructionally sensitive measurement instruments has been to gain and maintain effective cooperation between Curriculum and Evaluation. Some of the principles and practices that have helped us succeed in this crucial but delicate part of the work are the following:

- We established and followed the principle that Curriculum would have the leadership and the final say so in determining what was to be measured, while Evaluation would lead on issues of how to measure.

- We shared control and resources. For instance, in Portland we formed a Curriculum and Evaluation Council made up of the senior personnel from each area. This group, which meets monthly, is chaired in alternate years by the heads of the Curriculum and Evaluation Departments. In the meetings information is shared, policy direction is agreed upon, resources are allocated, plans are made and monitored, task forces are created for such work as test blueprint design and reports, revisions, and recommendations of the task forces are reviewed, revised, and ratified.

- We learned how to negotiate (and even enjoy) good, clean conflict since our ends were always the same (namely more informed and better decisions about students, programs, and policies), whereas our disputes were about means which themselves could usually be put to the test of data through experimentation and pilot testing.

- We took the long view. Blessed by enlightened top administrators and school boards, we were able to count on the resources and support to plan and develop on five, ten, and even twenty-year horizons.

Quality Control

Quality control measures are employed to insure that all aspects of the testing program are functioning accurately and effectively. Objectives of quality control and means taken to accomplish them are described below:

1. Insure that schools receive pre-printed answer sheets that are organized as requested by each school (by homeroom, by subject matter, by class, by grade level, etc.) for test administration.

   The school master file showing how each school requested its answer sheets to be organized in the next prior testing is compared with the current set of request forms received from principals. Changes in these requests are identified (usually only a few) and the master file is updated. This file automatically governs the way pre-printed answer sheets are ordered for distribution to the schools.

2. Insure that booklets that are distributed are printed properly, have all pages, etc.

   Samples from each printing order are taken to insure that all pages are present, all items appear in the test, and quality of printing is satisfactory.

3. Insure that the quality of printing on the pre-printed answer sheets will permit accurate scanning.

Pre-printed answer sheets for each test administration are "fan scanned" to find any answer sheets in which the quality of print may not permit accurate scanning, and that all needed information is in its proper place.

4. Insure that the computer program for pre-printing name, I.D., grade level, and test code on answer sheets is functioning accurately.

All programs for pre-printing answer sheets are test-run and results are examined to insure that data are accurate and complete.

5. Insure that students are assigned the proper level of a test series.

The program that assigns students to their proper level tests in any given administration is test-run, and the results for a sample of students are examined in relation to previous test results to insure that the program is functioning properly. (Placement is based on an algorithm averaging each student's five prior test standard scores with double weighting on the last standard score. This average standard score is converted to a RIT score and level placement for the grade level and time of year the testing is being carried out.)

6. Insure that the students understand directions for taking the test.

Verbatim instructions for administering the PALT are provided every teacher, and the testing procedures involve completion of sample items to assure student understanding of how to use the machine-scored answer shcets. There are specially designed answer sheets and practice tests for use with beginning third grade students.

7. Insure that conditions of test taking are good.

Meetings are held with test building coordinators prior to each administration and guidelines for conditions of test taking, as outlined in manuals of instruction for coordinators and school personnel, are addressed at these meetings.

8. Ascertain validity of student performance.

Both the coordinator and teacher manuals state criteria for labelling answer sheets as invalid where student performance does not meet these criteria. In addition computer programs label as invalid any performance falling outside the valid measuring range of the test form the student is assigned to. Also, if a student has 10 or fewer item responses (items attempted) on an answer sheet, the computer notes that performance as invalid.

9. Insure that answer sheets are properly marked and completed.

Each teacher must complete a signed checkoff that he/she has inspected each answer sheet for erasures, multiple marks, extraneous marks, and use of inappropriate markers on the header sheet before the sheets are returned. Scanner programs are used when tests are scored to identify incomplete information for essential hand-coded data.

10. Insure that coded information on each answer is correct and complete.

The procedure described under 3 above insures that pre-coded information is correct.

11. Insure that students have taken the assigned form of the test.

For all students scoring below chance levels, each test is rescored using the scoring keys for all forms in the series. This identifies any student whose test form is different from the one coded on the answer sheet. In such cases the legitimate score is entered on the student's record.

12. Insure that students needing to be retested are retested (high and low).

Students scoring outside the valid measurement range of the test are made known to the school and assigned a higher or lower test .form as appropriate (two levels higher or lower, if possible). Schools are provided the materials required for retesting and all schools routinely schedule such retesting following the regular program. Where retests are regarded as inappropriate by teachers, they are not required. In such cases low students receive no test record and high students retain the score achieved.

Complexity and Credibility

Making a complex system credible to the psychometrically unsophisticated is a major issue we have faced in developing IRT-based item banks and functional level testing systems based upon them. This becomes an especially critical matter when the test results seem to suggest something negative about the performance of students, programs, etc. Rhetoric such as "voodoo testing" is only a breath away when low levels or small gains are registered by complex measurement systems.

There is no magic answer to this problem. Its solution must be approached on a variety of fronts, such as a good inservice program; frequent helpful communication with representatives of the media and with community groups; effective collaboration with public information departments; as well as production and dissemination of a layered system of media productions to tell the measurement system's story--ranging from brochures through workshops and manuals to effective use of electronic media such as video cassettes and cable television. It also helps to periodically have the system evaluated by independent measurement experts who can both sanction the program and suggest improvements.

WHERE WE ARE GOING

New Testing Directions

Since the development of the item bank capability, we have found ourselves continually being asked by teachers, principals, and curriculum specialists for more and better measurement instruments and test results reports. They are developing creative and productive ways to use good data on how students are performing relative to the goals of the district. It is indeed a refreshing change to have the requests for new information come from teachers and principals as well as from the School Board or Superintendent.

For instance, this year 42% of our Black students are reading above the national average up from 36% two years ago and 31% four years ago, an eleven-point gain. Since the national average is the point above which 50% of students achieve, we are now within eight percentage points of having this student population at the national level and are gaining at a rate at which that goal should be reached within three years.

We have been working to develop a scale for measuring attitudes toward the subject of mathematics, and have been exploring ways to create scales for other affective variables. Affective goals are a part of many school districts' curriculum continuums, and yet very little has been done to provide information back to teachers and other instructional leaders as to how effective they have been in developing those traits. We recognize the potential that measurement of affective traits could have in the classroom, but we have only started to develop one such scale and to explore the possibilities for others.

Currently, NWEA and many school districts including Washington, Oregon, California, and Missouri have begun a massive project of developing the capability of measuring Science goals in grades K through 12. Unlike the basic skills item banks, Science has a very diverse curriculum across the participating districts. This diversity produces constraints on the scaling process as well as pointing out a tremendous need to utilize the flexibility of item bank-based measurement by individual school districts.

We are currently examining data gathered in a field testing designed to explore the dimensionality of the underlying traits (Kingsbury, 1985) typically taught in Science. This data will provide the structure to begin developing the scales for measurement. The five-year measurement objectives of the Science project are to develop as many as 16,000 items distributed across the Science goals. A particularly unique emphasis of the project is to focus not only on content-related items, but also on process and concept goals. Many concepts in Science transcend the particular content in which it is being taught (i.e., conservation, equilibrium cycle, etc.). Developing items to measure these concepts that are independent of the context of a particular subject area is meticulous and thus time consuming.

New Testing Technology

In an effort to allow teachers and school administrators to take advantage of the item banks and modern testing procedures, further applications of computers in the classroom are being explored. As an example, systems are currently being designed to allow the storage and retrieval of test questions from a video disk.

Since the storage capacity of a video disk is so great, this would allow the efficient use of very large item banks. In addition the capability of the video disk to incorporate motion and sound will allow the expansion of the current item banks to include simulation questions, progressive response questions, and questions which do not force the student to read. This will allow the expansion of the CAT program to the more accurate measurement of the basic skills and the expansion of our current measurement to higher level skills.

Another direction of development with the item banks is the technology to allow computerized test design and generation. This new capacity will allow district-wide tests to be developed and produced in a more efficient manner. In addition this technology could be used in the future by teachers and administrators to directly access the item banks to enhance school-based decision making. This use of the computer to store and maintain item content and statistics will make the problem of administering item banks somewhat more manageable.

A basic goal of the item bank work is to close the gap between testing and instruction. One phase of this is to produce tests which have more direct instructional consequence. The CAT pilot project is a first attempt in this direction. By developing tests which more efficiently test the basic skills, more testing time is left to perform pre-diagnostic and diagnostic testing. The ultimate goal of this research effort is to create a more viable interface between the teacher and the evaluator (in this case in the form of the computer).

HOW OTHERS CAN SHARE

Readers who have come this far may be wondering how they can gain access to the item banks, tests, and technology we have been describing and, perhaps, how they can participate in their further development. Here are some answers:

- The current Reading, Mathematics, and Language Usage banks are being disseminated to local school districts for their own use by the Northwest Evaluation Association under a license from the Portland School District. Contact can be made by writing:

Ray Miller, Executive Secretary
NWEA
ESD #121
1410 South 200th
Seattle, WA 98148

or by phoning: 206-242-9400, Ext. 58, or 206-839-3932.

Participation in the ongoing collaboration within the NWEA framework can be initiated through the same source.

- Licenses for use of the current three basic skills item banks by others than local school districts may be requested from the Portland School District by contacting:

Dr. Walter E. Hathaway, Director
Research and Evaluation Department
Portland Public Schools
501 N. Dixon Street
Portland, OR 97227
(503) 249-2000, Ext. 206

- Information concerning participation in the NWEA Science Curriculum and Assessment project may be obtained by contacting:

Susan Smoyer
NWEA Science Project
700 Pring]e Parkway S.E.
Salem, OR 97310
(503) 378-4157

- Questions concerning the Computerized Adaptive Testing project and other applications of mode..n technology to testing should be addressed to:

Dr. G. Gage Kingsbury
Research and Evaluation Department
Portland Public Schools
5C1 N. Dixon Street
Portland, OR 97227
(5)3) 249-2000, Ext. 229

- Information about the Portland Achievement Levels Tests and the application of IRT to the item banks may be obtained from:

Dr. Ronald L. Houser
Reseaıch and Evaluation Department
Portland Public Schools
501 N. Dixon Street
Portland, OR 97227
(503) 249-2000, Ext. 253

- The Portland Achievement Levels Tests and the scoring and reporting software are published and marketed by Microprocessors for Education and Business, 109 N. Main Street, Gresham, OR 97030 (503-666-7883).

- The results of the past and ongoing research and development related to our efforts are available in the form of reports, monographs, conference presentations, articles in popular and refereed journals, book chapters and books, as well as on electronic networks including the American Educational Research Association's COMPUSERV system and The Council of Great City Schools' Telenet.

We welcome the interest and involvement of our fellow researchers, developers, and educators in this important work.

References

Gardner, D. P. (Chair) (1983). A Nation at Risk: The Imperative for Educational Reform. Washington, D.C.: U.S. Government Printing Office.

Haney, W. (1985). Making Testing More Educational. Educational Leadership, October, 4-13.

Hathaway, W. E. (1980). A School-Discrict-Developed, Rasch-Based Approach to Minimum Competency Achievement Testing. In R. M. Jaeger & C. K. Tittle (Eds.) Minimum Competency Achievement Testing: Motives, Models, Measures, and Consequences. Beverly Hills, CA: McCutchan.

Hathaway, W. E. (Ed.) (1983). Testing in the Schools: New Directions for Testing and Measurement, Number 19. San Francisco, CA: Jossey-Bass.

Houser, R. L., Hathaway, W. E., & Ingebo, G. S. (1983). An Alternate Procedure to Obtain Ability Estimates in Latent Trait Models. Paper presented to the annual meeting of the American Educational Research Association, Montreal, Canada.

Kingsbury, G. G. (1985). A Comparison of Item Response Theory Procedures for Assessing Response Dimensionality. Paper presented to the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1983). Small N Justifies Rasch Model. In Weiss, D. J. (Ed.) New Horizons in Testing: Latent T ait Test Theory and Computerized Adaptive Testing. New York: Academic Press.

Lord, F. M. & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

Swaminathan, H. & Gifford, J. A. (1985). Bayesian estimation in the two-parameter model. Psychometrika, 50, 349-364.

Weiss, D. J. & Kingsbury, G. G. (1984). Applications of Computerized Adaptive Testing to Instructional Problems. Journal of Educational Measurement, 21, 361-375.

0004E
11/21/85

IV. HOW CLOSE ARE WE TO REALIZING THE POTENTIAL OF ITEM BANKS

Gary D. Estes
Northwest Regional Educational Laboratory

To realize the potential of item banking we need to minimize "reinventing the wheel" in developing local tests. The Wisconsin and Portland item banking systems were selected for this report in part because they are useful for outside organizations and individuals. One objective in NWREL's Assessment Development and Use Project was to explore areas in which work done by other agencies or systems could be shared, and to reduce the effort needed by others to undertake similar projects. Portland and Wisconsin systems illustrate major advances in realizing the potential of item banks.

In reacting to the Portland and Wisconsin systems and commenting on the question of the potential of item banks, I would like to offer comments in relation to issues outlined by Arter and Estes (1984) in how item banks work and how well these item banks serve the purposes for which item banks are developed.

Prior to discussing item banking issues or purposes, it might be helpful to review the definition used for item banks by various agencies to illustrate that we differ in our views of item banks and what are their key critical characteristics. Portland's definition differentiates an "item bank" and "item pool" as a function of whether the items have statistical characteristics such as Rasch calibrations. Wisconsin described an item bank according to three variables and inferred that item banks could fall on a continuum in this three-dimensional space. Estes and Arter (1984) loosely defined an item bank..."as a large collection of distinguishable test items." We elaborated that "collection" meant that the items were kept together in some retrievable form; "distinguishable" meant that the items carried information that allowed for the selection of items for tests, and that "large" meant that the number of items would be greater than that used in any one testing. The purpose in reviewing our definitions of item bank is to highlight that it is not always the case that a common perception exists in using the term. The objective for mentioning it now is simply to recognize the differences rather than to debate which definition would be most useful or under what context and purposes different definitions would be relatively more or less appropriate. My view is "if the definition fits, use it."

Issues Related to How Item Banks Work

Millman and Arter (1984) provided a list of issues in item banking. They classified issues into those dealing with items, tests, system, use and acceptance, and cost categories. I would like to organize my comments around these issues.

Item Issues

The Wisconsin and Portland examples both represent cases in which existing item collections were used in developing their item banks. Wisconsin in fact made extensive use of the Northwest Evaluation Association (NWEA) item bank which was developed largely through Portland's efforts and support. Both related that they found categories in which items were plentiful and areas in which they needed to develop or revise items. Their experiences match those with our earlier surveys in which items in the areas of reading, math and language arts appear to be relatively plentiful for the elementary school years, but increasingly sparce in the high school or other subject areas. Evidence of this is the degree to which the NWEA is needing to write items for their science project. The development of new items for this project is significantly greater than that in the earlier reading, math and language arts efforts.

A generalization from these two case studies and other evidence is that many sources of items exist in relatively few subject areas. However, there are increasing numbers of items in other areas such as science, high school subjects and affective domain categories.

Classification of Items. This area represents one of the more interesting issues in item bank development and use. It is clear that one of the primary reasons for developing tests locally is the need to match assessments with local objectives and curriculum. The local matching is an area in which item banking can be most helpful. We would hope that given ten variations for a math schema or scope and sequence in grades 2-8, that a new test developer or "item banker" could select one that would be very close to meetinq their needs. That does not happen in reality. Wisconsin's experience is fairly typical in that even with large and broad representation across the state in developing schema there is still distortion as districts begin to use the item bank schema in matching their locally established competencies.

A guideline is that anyone undertaking the development and implementation of a future item banking effort should anticipate substantial work in item classification systems. One view might be that this is negative because insufficient use is made of existing classification schemes. An alternative, more positive view, is that for item banks to adequately capitalize on the perceived and actual need to match local objectives, it will be necessary for each local agency to go through a process that results in a classification scheme in which that agency or group has ownership. appears that until sufficient planning and conscience decision making has been invested, it is will be improbable that the local identification and ownership will result.

Managing Items. A question which often arises is the extent to which items will be revised or systems will allow for revision and updating of items. Portland's strong commitment and dependence on the item calibrations requires that all items that are revised go through the pilot testing and quality control checks. Wisconsin also allows for item revisions. But their lack of a strong item information data base results in their primary item revision concerns being item formatting, entry and logistical management issues. These item revision/management issues should not be underestimated. Users will want to revise items and these revisions will need to be accomodated. For example, Wisconsin decided that a single rather than a combined horizontal and vertical format for math items was desired. One of the first districts to use the Wisconsin item bank determined that, "We do not present math problems in that format." Thus, for their tests they reformatted items into a vertical format. If item calibration or difficulty information were available, it would not be valid for the reformatted items, i.e., vertical and horizontal math problems do not have equal difficulty.

An implication is that probably regardless of the item bank size, individual users will invariably identify alternative item formats or versions that will be preferable or will be an "improvement" over the existing item(s). As Wisconsin, and to a lesser extent Portland, have related, when to allow revisions and whether to enter the revised items into an item bank are key nontrivial issues. It would be relatively easy to have a large number of somewhat minor variations diminish the utility of an item bank. If users have to make fine discrimination in item variations rather than helping them to develop a test matched well to their specifications, this will become a liability.

Item Maintenance. The storage and management of items differ between Wisconsin and Portland. Both allow for use of items without reentry and typing each time an item is used. Portland maintains a "camera-ready" hard copy of items that can allow for item use without retyping or entering. Wisconsin's graphic and text capabilities allows for retrieving and managing items in an electronic file. Both systems are moving to more fully automate the management and retrieval of items. An item bank that enables items to be used without retyping, editing, etc., gains a great advantage of item banking's potential.

Test Issues

The Wisconsin and Portland systems differ in the way in which tests are assembled. A major difference is that Wisconsin's items are stored in an electronic data base whereas Portland's are still maintained in a hard copy format. Although Wisconsin's test assembly is an advantage to districts close to the Wisconsin State Department, it would be very difficult to assemble tests from the Wisconsin item bank (or to even share the Wisconsin item bank) outside Wisconsin. Thus, the hard copy format for the NWEA and Portland item bank, while not offering a "streamlined test assembly potential," does offer a test assembly potential for users. A simple purchase of or access to the Portland item bank allows for item selection and test assembly in a manual but locally controlled system. As outlined in the Wisconsin paper, further development such as computerized selection of items rather than manually selecting items from electronic files will enhance the test assembly process. Given the rapid advances in the technology and the further development in agencies such as Wisconsin, we can hope that test assembly will be more fully automated in the very near future. Deck and Estes (1984), Deck, Nickel and Estes (1985) have reviewed several imcrocomputer programs that can assist with item selection and test assembly.

The NWEA and Portland item banks do not have test assembly and printing as an inherent part of the item bank. Wisconsin's item bank test assembly and printing functions results in high quality tests. (For example, the Wisconsin paper was developed on the same Xerox system that maintains their item bank and prints tests.) A major difference between the Portland and Wisconsin items banks is the degree to which item information is maintained and can be used in selecting tests. As evidenced by Portland's movement toward computer-adaptative testing, item data such as calibrations provide flexibility in assemblying tests matched to either individual student abilities or desired test level difficulties.

Neither system currently supports test administration, scoring and reporting in the item bank. However, Portland's CAT testing is moving to integrate these functions. A primary objective within the Portland system is to move the test administration, scoring and reporting to the classroom level and to obtain school and district testing results from classroom level information.

System Issues

It is clear that in both cases the development of these item banking systems required substantial fiscal and personnel resources. The commitments of top-level policy makers over a long-term period was largely responsible for the support to develop these systems. Although both systems have fairly large user audiences, neither initially was undertaken with the purpose to transport the system widely. The NWEA and Portland item bank has, however, been widely distributed throughout the Northwest states and districts. Thus, the resources from the NWEA item bank are available to others for a fraction of the cost required to develop a similar system. Others, however, will likely need moderate to major efforts to adapt the NWEA or other item banks to their own purposes. Support for this conclusion is the degree to which the development of item classification schemes, the review and revision of items was needed when Wisconsin used the NWEA and other item banks.

The hardware and software features of the item banks in Wisconsin and Portland serve different functions. Portland has an extensive system for managing information about items. Their data base system enables Portland to calibrate and maintain item use statistics on their item bank. The software and hardware in Wisconsin provides flexibility and power in formating items and tests. The two computer system plan outlined in Wisconsin's paper has not yet been realized. It is likely that in Wisconsin the management of item information will be slow in development given the current stage of the system and variable fiscal and political support for further development.

One conclusion is that although graphics, text and data might all be managed within a single system, there is none operational that can be widely shared with school districts or state departments. Wisconsin's experience offers hope that many of these problems can be resolved with decreased costs for hardware and increased software development. If a district or other agency are considering an item bank, they should apply as much of the development in systems such as Portland's and Wisconsin's as fits their fiscal and test development resources.

Monitoring and Training. Both systems require very little training to use the item banks to develop tests locally. Wisconsin provides ongoing training in several areas to support not only the use of the item bank, but to improve local district's test development and use skills. Portland and the NWEA have also provided streamlined training procedures to enable people to use the item bank and to do "sophisticated" technical procedures such as calibrating items that previously required much training and measurement background. Thus, these two examples offer evidence that making good and efficient use of an item bank does not require extensive training.

Use and Acceptance

The NWEA item bank has been in use for some years and has widespread acceptance in major districts and student population areas within Washington and Oregon. Several areas outside Oregon, e.g., Wisconsin, have also made extensive use of the NWEA item bank. There is much evidence through the number of districts and strong commitments to the NWEA item bank efforts of its acceptance within these agencies. This probably derives largely from the districts' sense of involvement and ownership with the item bank.

The degree to which the Wisconsin item bank will be used and accepted is not yet fully determined. However, in talking with districts that have made use of the item bank, they were very enthusiastic about the degree to which the item bank was able to support their local test development efforts. Although each district without exception expressed needs to revise items and tailor them to their own needs and local curriculum, each expressed that they believed that they had higher quality items for less effort than otherwise would have been possible.

Cost

It is clear that both of these items banks required large investments of both fiscal and personnel resources. The Wisconsin item bank hardware and software cost in excess of $200,000 with over $30,000 per year needed for system maintenance. The number of staff needed for the Wisconsin item bank has ranged from 12 in 1983 to seven during the most recent year. This could be viewed as enormous costs if it had to be subsumed within a single district. However, the costs become much more modest when examined on a cost per district or per pupil basis.

Similarly, the developmental costs for the NWEA item bank could not be borne solely by the Portland Public Schools. The NWEA cooperative approach in developing the science project represents an approach to large item bank development efforts.

Cost might be significantly reduced for those who wish to build on resources such as those in the NWEA and Wisconsin item banks. Limitations on the potential cost savings will be directly proportional to the degree to which another agency needs to tailor the item bank and system to their own needs. Costs associated with developing classifications systems, entering items into the system, and ensuring items meet the item bank and item specifications, will increase costs beyond simply purchasing or obtaining an item bank and using it. As the Wisconsin paper outlined, there will be some distortion when another system is used for one's local test development. The trade-offs between adopting or adapting another's system is one of the key issues in item banking.

There is not a simple response to the question or whether costs are prohibitive. The proliferation of item banks and item pools around the country gives testimony of the degree to which local ownership and development is at least implicitly judged to be worth the investment of resources. Several agencies, of which Portland and Wisconsin are prime examples, have made significant efforts to share their developments with others.

In summary, the costs associated with building an item bank from existing resources such as Portland's and Wisconsin's will be less than proceeding independently. Whether these costs are justifiable depend on (a) the demand for locally developed tests, (b) financial resources available and (c) the number of tests likely to be needed.

How Well Do Item Banks Fulfill Their Potential

Estes and Arter (1984) outlined five areas in which item banks would have the most benefit: (a) providing a match to local objectives and curriculum; (b, enabling frequent testing to occur; (c) allowing for individually tailored tests, e.g., the adaptive testing described in Portland's paper; (d) developing multiple equivalent test forms for secure or repeated testing; and (e) providing support with more difficult items, e.g., reading passages, graphs. Does the experience and information from Wisconsin, Portland and others support the potential of item banking for these purposes?

All item banks we have surveyed (Estes and Arter, 1984) were generated to meet a primary need to have assessments matched to local objectives. User flexibility in adapting the systems and items to their local needs has promoted the ability of these and other item banks to serve this purpose. Again, the potential value of item banking is qualified by the effort needed to make adaptations to ensure that the local match is achieved.

The frequent testing advantage is illustrated somewhat differently in the Wisconsin and Portland examples. The item bank in Wisconsin is designed and will be capable of serving multiple district test development efforts. Similarly, Portland has served a wide range and number of districts. Thus, both systems have met the purpose of serving frequent testing needs where frequency is defined by the number of district users.

However, Portland is also moving to a frequent <u>student testing</u> mode through their computer-adaptative system. The computer-adaptive testing system in Portland is dependent upon the item response theory approach to developing their item banking system. While this system may not be critical for all purposes, e.g., classroom teacher use of tests, item calibrations serve a valuable purpose and are enabling Portland to accomplish multiple testing purposes from their item bank. As outlined in Wisconsin's paper, they also hope to establish better item information and have begun to develop and calibrate items.

The Portland system has for several years used functional level testing and thus has served one need for student tailored tests for a period of time. This functional level has not place as high a priority on individually tailored tests as the new computer-assisted testing does. It should be reemphasized that the individually tailored testing is possible with the item calibrations that were emphasized throughout the Portland paper.

Multiple equivalent forms could be developed from either item bank. However, Portland's would have the advantage of having equivalence determined on the basis of both content and comparable difficulty levels.

User's of both the Wisconsin and Portland item banks (as well as others) have expressed strong support for the value of item banks in supplying good, well developed reading passages, charts/figures and even item formats. At a minimum, users have been able to write new items much more efficiently than would have been possible without these. Thus, item banks do assist in this area and result in lower costs and potentially better quality items.

Summary

If there is a relatively high demand, strong political and administrative support, and some fiscal and staff resources, item banks can be a valuable resource in local test development. Many of the characteristics and information outlined in the Portland and Wisconsin item banks will be helpful. It will not be necessary, however, to have the same priorities placed on certain characteristics, e.g., the item response theory based item bank in Portland or the sophisticated item format and graphic capabilities in Wisconsin. One will, on the other hand, need to invest resources into adapting and revising other's item classification system and developing decision rules for allowing/accommodating item revisions.

Our hope is that other's will continue to share lessons learned as they build on existing item banks such as those described here. This cooperative spirit will help to insure that sound and useful local assessments are conducted to support effective and accountable education.

References

Estes, G., & Arter, J. (1984). *Item Banking for State and Local Test Development and Use*. Portland, Oregon: Northwest Regional Educational Laboratory.

Estes, G., & Arter, J. (1984). *A Guide to Item Banking in Education*. Portland, Oregon: Northwest Regional Educational Laboratory.

Millman, J., & Arter, J. (1984). "Issues in Item Banking." *Journal of Educational Measurement*, *21* (4):315-330.