AFHRL-TP-88-40

(2)

# AIR FORCE HUMAN RESOURCES LABORATORY

AD-A205 370

## AUTOMATED ITEM BANKING AND TEST DEVELOPMENT



William M. Lee
Pamla Palmer

Operational Technologies Corporation
5825 Callaghan Road, Suite 225
San Antonio, Texas 78228

Linda T. Curran

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

February 1989

Final Technical Paper for Period October 1987 - April 1988



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

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

# REPORT DOCUMENTATION PAGE

Form Approved OMB No 0704-0188

| | |
|---|---|
| 1a. REPORT SECURITY CLASSIFICATION<br>Unclassified | 1b RESTRICTIVE MARKINGS |
| 2a. SECURITY CLASSIFICATION AUTHORITY | 3 DISTRIBUTION/AVAILABILITY OF REPORT<br>Approved for public release; distribution is unlimited. |
| 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S) | 5 MONITORING ORGANIZATION REPORT NUMBER(S)<br>AFHRL-TP-88-40 |
| 6a. NAME OF PERFORMING ORGANIZATION<br>Operational Technologies Corporation<br>6b OFFICE SYMBOL (If applicable) | 7a. NAME OF MONITORING ORGANIZATION<br>Manpower and Personnel Division |
| 6c. ADDRESS (City, State, and ZIP Code)<br>5825 Callaghan Road, Suite 225<br>San Antonio, Texas 78228 | 7b. ADDRESS (City, State, and ZIP Code)<br>Air Force Human Resources Laboratory<br>Brooks Air Force Base, Texas 78235-5601 |
| 8a. NAME OF FUNDING/SPONSORING ORGANIZATION<br>Air Force Human Resources Laboratory<br>8b OFFICE SYMBOL (If applicable)<br>HQ AFHRL | 9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER<br>F41689-87-D-0012 |
| 8c. ADDRESS (City, State, and ZIP Code)<br>Brooks Air Force Base, Texas 78235-5601 | 10 SOURCE OF FUNDING NUMBERS<br>PROGRAM ELEMENT NO: 63227F; PROJECT NO.: 2922; TASK NO: 02; WORK UNIT ACCESSION NO: 02 |
| 11. TITLE (Include Security Classification)<br>Automated Item Banking and Test Development | |
| 12. PERSONAL AUTHOR(S)<br>Lee, W.M.; Palmer, P.; Curran, L.T. | |
| 13a. TYPE OF REPORT<br>Final | 13b. TIME COVERED<br>FROM Oct 87 TO Apr 88 |
| 14. DATE OF REPORT (Year, Month, Day)<br>February 1989 | 15. PAGE COUNT<br>38 |
| 16. SUPPLEMENTARY NOTATION | |

| 17. COSATI CODES | | |
|---|---|---|
| FIELD | GROUP | SUB-GROUP |
| 05 | 08 | |
| 05 | 09 | |

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
computer systems; item storage; item analysis; pilot selection; item bank

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Projects to develop an automated item banking and test development system have been undertaken on several occasions at the Air Force Human Resources Laboratory (AFHRL) throughout the past 10 years. Such a system permits the construction of tests in far less time and with a higher degree of accuracy than earlier test construction procedures. This paper details Classical Item Theory and Item Response Theory (IRT) approaches to item banking and test construction and their relevance to the development of an automated item banking system. State-of-the-art improvements to the current automated item banking system are proposed which include the capability to generate multiple forms simultaneously and to print new test forms with the same type font, spacing, and format as the reference form.

| | |
|---|---|
| 20 DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS | 21 ABSTRACT SECURITY CLASSIFICATION<br>Unclassified |
| 22a. NAME OF RESPONSIBLE INDIVIDUAL<br>Nancy J. Allin, Chief, STINFO Branch | 22b TELEPHONE (Include Area Code)<br>(512) 536-3877<br>22c OFFICE SYMBOL<br>AFHRL/SCV |

DD Form 1473, JUN 86 *Previous editions are obsolete.*

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

AFHRL Technical Paper 88-40 February 1989

# AUTOMATED ITEM BANKING AND TEST DEVELOPMENT

William M. Lee
Pamla Palmer

Operational Technologies Corporation
5825 Callaghan Road, Suite 22[illegible]
San Antonio, Texas 78228

Linda T. Curran

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

Reviewed by

Linda T. Curran, Acting Chief
Enlisted Selection and Classification Function

Submitted for publication by

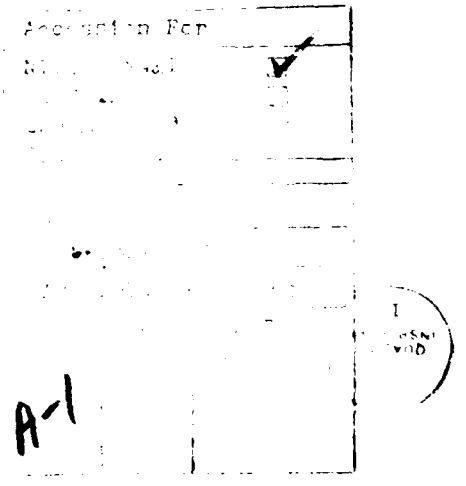
Lonnie D. Valentine, Jr.
Chief, Force Acquisition Branch

## SUMMARY

The Armed Services periodically require the development of new forms of the Armed Services Vocational Aptitude Battery (ASVAB), the selection and classification instrument used to qualify individuals for military enlistment. These new forms of the ASVAB must be parallel to one another and to a reference test (ASVAB Form 8a). In addition, when printed, the new forms must have the same type font, spacing, and format as the reference form. Any deviation may result in their not being parallel to the reference form. With the advent of high speed computers which offer increased speed and flexibility, the Air Force Human Resources Laboratory (AFHRL) has investigated ways to develop new forms more efficiently and precisely. Over the past 10 years, AFHRL projects have undertaken the development of an automated item banking and test development system. This paper describes these previous efforts and current concerns that indicate revisions are necessary.

The approach to design of an automated item banking and test development system can be envisioned as an integration of three phases. Phase I involves the development of an item bank that contains each item's content (i.e., text) and statistics and has full editing capabilities. Phase II provides the capability to retrieve items from the item bank and to construct tests with prespecified characteristics. Finally, in Phase III, the actual test booklets are published with a specific type font, spacing, and format.

Recent concerns, as described in this paper, have indicated that an update of the current automated item banking and test development system is necessary to obtain a fully integrated system. Previous efforts at developing an automated test development procedure have not resulted in a comprehensive collection of ASVAB items (Phase I). A concentrated effort to bank previous ASVAB items' content and statistics has been undertaken in the present effort because of recent policy changes allowing a percentage of previous items for reuse in new forms, and because of technological advances in printing from computer-based data banks. The focus of previous efforts in developing a system has been on the test development phase, Phase II; however, the currently implemented system can develop only one test form at a time. An improvement for this phase would be to include the capability to develop more than one parallel form simultaneously. Lastly, Phase III, the publishing phase, has not been addressed by previous efforts. Enhancement to the current system would make it possible to print tests having the same type font, spacing, and format as the reference form.

## PREFACE

The work documented in this technical paper was completed as part of the Development of Armed Services Vocational Aptitude Battery (ASVAB) Forms 20, 21, and 22 (Items and Item Bank). The paper was prepared by Operational Technologies Corporation, San Antonio, Texas, under contract F41689-87-D-0012. Work was conducted under work unit 29220202 for the Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB, Texas. The authors would like to thank Carl Haywood of Operational Technologies Corporation for his valuable assistance in the preparation of this report.

## TABLE OF CONTENTS



## LIST OF FIGURES

# I. INTRODUCTION

In order to provide the best quality enlisted force, the armed services periodically require the development of new selection and classification tests. To meet the testing requirement, the military services use the Armed Services Vocational Aptitude Battery (ASVAB). The proper allocation of manpower resources can result in a substantial benefit because of reduced attrition and higher productivity as well as increased personnel morale and greater reenlistment rates.

The ASVAB is a multiple-aptitude test battery used by all of the armed services for selection and classification. This test is administered to over 1 million applicants yearly in about 69 Military Entrance Processing Stations (MEPS) and approximately 900 Mobile Examining Team Sites (METS), and to about 1.3 million high school students yearly in over 14,000 high schools across the nation. In order to reduce the possibility of compromise of the ASVAB, periodic development of new forms is required.

The advent of high-speed computers has led to the possibility of constructing tests of known quality on computers (Ree, 1978). This process or system of developing computer-generated tests can be viewed as involving 3 phases. The first phase includes the development of an item bank that contains the content and statistics of a large pool of items and has full editing capabilities. Phase II provides the capability to retrieve items from the item bank and to construct tests with prespecified characteristics. Phase III involves the automated publishing of these tests with particular font, spacing, and format specifications. Ree established that the use of such a system can improve the security of stored test items, and so can provide protection against loss or compromise.

## A. ASVAB Automated Item Banking and Test Construction

As stated, Phase I is the banking of the content and statistics of items with known qualities or parameters. The parameters include but are not limited to item statistics, such as item difficulty and item discrimination, and statistics of the test where the item appeared, such as test mean and standard deviation. The parameters of the items can then be used to retrieve items in order to construct tests with prespecified characteristics (Phase II). In order to provide Phases I and II of the system, Ree (1978) developed an Automated Item Banking (AIB) system that permitted the banking (storing) of items on a Sperry-Univac mainframe computer and the construction of tests with known parameters while the user (test constructor) is on-line. (Further elaboration of Ree's AIB system will be provided in further discussion.) An advantage of Ree's AIB system is that it is simple enough to use without detailed knowledge of computer programming or operation. Furthermore, the speed and flexibility of the AIB system permitted the construction of tests in far less

time and with a higher degree of accuracy than earlier test construction procedures (see Thorndike, 1971, for a discussion of test construction issues).

### B. The Need for Parallel Forms

In testing situations where the ASVAB is used, it is necessary to convert or relate test scores obtained on one test to those obtained on another. For this type of horizontal equating to be accurate, the tests must have comparable levels of difficulty; and be parallel (in terms of content and raw scores) to one another and to a reference form. (For ASVAB, the reference form is currently 8a.) Constructing such parallel forms involves the assembly of items into tests to meet diverse specifications simultaneously. Research on the development of improved technologies for selecting test items from an item bank and for assigning them to tests in a way that maximizes the psychometric similarity and general merit of the resultant test forms was conducted by Lee and Fairbank (1983). Their purpose was to refine the existing item banking system (AIB system) to make it more effective and efficient in the process of test construction. However, the refinement of the AIB system does not offer the flexibility of generating multiple parallel forms simultaneously.

### C. Present Effort

This paper specifies recommendations to improve or enhance the most recent version of the AIB system by taking advantage of recent advances in psychometric theory, mathematical modeling, and computer software. Before addressing current concerns, this paper will review classical test theory and item response theory (IRT) approaches to item banking and test construction, the role of mathematical modeling in the construction of parallel forms, and prior efforts in the development of the AIB system. Current concerns to be discussed include the need for a comprehensive item bank (Phase I); the capability to generate multiple forms simultaneously (Phase II); and the ability to print forms in the same type font, pitch, spacing, and format as the reference test (Phase III).

## II. REVIEW OF APPROACHES TO AUTOMATED TEST CONSTRUCTION

This section reviews classical versus IRT test construction, recent developments in the use of quantitative modeling and mathematical programming to construct simultaneously parallel tests, as well as prior efforts at the Air Force Human Resources Laboratory (AFHRL) concerned with the development of an automated item bank.

### A. Classical Versus IRT Automated Test Construction

In order to clarify the similarities and differences between

classical test theory and IRT in automated test construction, it is appropriate to discuss first the theoretical assumptions of these two models.

## 1. Classical Test Theory

Classical test theory is a model that represents the way in which errors of measurement influence observed scores, and considers the resulting effects of such errors on reliability, validity, and other quantitative aspects of test efficacy. The reader should keep in mind as classical test theory is discussed that the statistical characteristics of the total test depend entirely upon the statistical characteristics of the items used to build it, which in turn vary according to several factors, such as the ability of the group tested, the group's heterogeneity, and the test's length, all of which may vary from one occasion to the next. This lack of invariance is one of the leading criticisms of classical test theory.

In classical test theory, a test can be a measure of a single trait (unidimensional) or of a number of traits (multidimensional). The test item is the unit building block from which the composite test is constructed. An individual's observed score on a test is usually defined as the number of items the individual answers correctly.

The classical test theory model is described by its assumptions; if these assumptions are met (as in any model), then the results derived from the model are acceptable. Conversely, if the assumptions are not met, then the conclusions derived from the use of the model are questionable. Specifically, there are seven basic assumptions of classical test theory (Allen & Yen, 1979; Gulliksen, 1950):

a. Classical test theory is considered an additive model, with an examinee's observed score being equal to the individual's stable true score or true ability plus a certain amount of random error. There are times when an individual will answer correctly an item he does not know and will answer incorrectly an item he does know. This is reflected in the random error score and can be due to various factors, for example, fatigue or guessing.

The assumption of an individual's observed score being equal to the sum of the individual's true score and random error score can be extended to show that the variance of the observed score is equal to the sum of the true score variance and the error variance.

b. The observed score does not necessarily equal the true score, but if it were possible for an individual to take a test an infinite number of times without changing that individual's true score (e.g., without practice and fatigue effects), his or her mean (average) observed score would equal his or her mean

true score. This is assuming that the average random error score is zero.

c. The coefficient of correlation between true and error scores is assumed to be zero; because the error scores are random and not systematic, there is no reason to expect large errors to occur more often for persons with low true scores than persons with high true scores, or vice versa.

d. Test forms are parallel when they are equivalent in terms of content, observed score means, variances, skewness, kurtosis, and reliabilities.

e. The coefficient of correlation between the error scores on one test and the error scores on another parallel test is assumed to be zero.

f. If the observed score is used to predict the score that a person will make on a criterion measure, the coefficient of correlation of error scores with the criterion scores is assumed to be zero.

g. Test forms are true score equivalent if their true scores are the same when a constant value has been added to examinees' scores.

Given these assumptions, test developers use the following item and test statistics that are based on the classical test theory model: item difficulty, or p-value, which is the proportion of the total number of examinees who choose the correct response (the higher the p-value, the easier the item); item discrimination which is the correlation between item and total test scores; the mean, standard deviation, skewness, and kurtosis of all examinees' number-right scores; and reliability of test scores, or coefficient alpha, which is an index of precision of measurement. These item and test statistics are dependent on the population of examinees who take the test. Therefore, test item selection that is based on p-values and item discrimination values is meaningful for the construction of tests only for the sample of examinees on which the values were calculated, and for the population from which the sample was drawn.

Classical test theory models have been developed and used over a period of many years. Even though the use of classical test models is prevalent in test development, there exist many problems in applying these models in test construction. As previously mentioned, the item difficulty and discrimination indices are dependent on the specific samples for which they are calculated. For example, an item's p-value will be higher when the item is administered to a sample whose ability is higher than the average ability level of the population for which the item was intended. Thus, p-values are good measures of item

difficulty only for the sample intended.

In addition, the item discrimination index is dependent on the homogeneity of the ability levels within a sample as well as the homogeneity of the item content of a test. That is, item discrimination values will be higher when calculated from a sample that is heterogeneous in ability than from a sample that is homogeneous in ability. This outcome may be attributed to the established effect of group heterogeneity (the variance of the test scores) on correlation coefficients (Hambleton & Swaminathan, 1985; Lord & Novick, 1968).

Also, the greater the true score variance in a sample, the greater is the ability of an item to account for ability variance. The test score reliability correlation coefficient calculation is based on test score variability, which in turn is also dependent on the sample of examinees. It logically follows that the mean, standard deviation, skewness, and kurtosis of the score distributions will vary according to the ability levels of the specific sample and according to test content. Thus, classical item statistics have utility in item selection during test development only if the experimental sample of examinees is very similar to the sample on which the statistics will be applied.

The on-line test developer would need to be able to retrieve much more information than the classical item statistic values. Population or sample demographics, as well as test form characteristics, would need to be easily identified and considered in the test development process.

Another related problem in classical test models concerns the administration of parallel test forms. First, parallel forms reliability is difficult to achieve with most data sets since examinees may not obtain the same ability score on the second parallel test administration. This result could be due to several reasons, such as fatigue or motivation/anxiety effects. Therefore, classical test models usually yield underestimates of the alternate parallel forms reliability correlation coefficient. The second problem encountered with classical test models and parallel forms is that certain ability groups perform more consistently on tests than do other ability groups. For example, high-ability groups perform more consistently than do medium-ability groups (Hambleton & Swaminathan, 1985).

In summary, classical test theory yields indices that are dependent on the population of examinees who take the test and are useful only when the same items are administered to the same or equivalent samples or when strictly parallel test forms are administered to the same sample. Due to the constraints of this situation, test developers have been interested in a more workable theory that would resolve the previously mentioned shortcomings. One theory that has received considerable

attention is IRT.

## 2. Item Response Theory (IRT)

As in classical test theory, IRT is based on the notion that 'latent traits' or underlying characteristics or abilities are estimated from observed scores on a set of test items (Lord & Novick, 1968). In IRT, probabilistic models are used to specify the mathematical relationship between observable test performance (test score) and the measure of the unobservable latent trait. The various IRT models differ primarily in the number of item parameters (difficulty, discrimination, guessing) included in the model and in the item scoring procedure. The items may be scored right/wrong, as are most standardized achievement test items, or they may be scored by awarding different numbers of points for varying degrees of correctness, as are many mathematical problem solving items. The models most frequently used with dichotomously scored items are the one-, two-, and three-parameter logistic models. The models most frequently used with polychotomously scored items are the nominal, graded response, partial credit, and continuous logistic models. Because this effort will be restricted to dichotomously scored multiple-choice items, only the former group of models will be discussed in detail.

As with any test theory, there are assumptions about an individual's performance on a test. Four assumptions and/or properties of the IRT models are as follows:

a. The first assumption is that the test is unidimensional--that is, the items on the test measure a single trait. This is a stricter assumption than is made by classical test theory, which encompasses tests made up of both unidimensional and multidimensional sets of items. Unidimensionality is a property of the items and is not affected by the ability distribution of the group of persons tested. The unidimensionality assumption does not imply that all items must correlate positively with each other. Items may correlate negatively with each other and still be unidimensional. It should be noted that multidimensional IRT models exist (see Reckase & McKinley, 1983; Samejima, 1974; Sympson, 1978); however, the most commonly used IRT models are ones that assume unidimensionality.

b. Historically, the second assumption is usually stated as that of local independence. The weak form of the local independence assumption states that for persons of the same ability level, item scores are uncorrelated; in other words, a person's response to any one item on a test is not rectilinearly related to their responses to any of the other items on the test. The strong form of the local independence assumption states that a person's item responses are statistically independent, which means that there is neither a rectilinear nor a curvilinear

relationship between the item scores of people with the same ability. In either case of local independence (strong or weak), the joint distribution of the item scores is considered to be equal to the product of the marginal probabilities--that is, the probability of answering all the items correctly is equal to the product of the separate probabilities of correctly answering each of the items. However, Lord (1980) explains that it is not an additional assumption, but follows directly from unidimensionality. Both forms of local independence will be satisfied if all of the test items measure a single ability. An important point is that local independence is conditional on ability level and in no way suggests that item scores are unrelated to each other for the total group of examinees.

c. In theory, item response models utilize parameters that can be estimated in a reasonably precise manner from item responses obtained from any sample of persons, whether or not the sample is representative of the population as a whole. Classical test theory, on the other hand, utilizes item parameters that can be estimated reasonably precisely only from a representative sample. Also, item response models theoretically can provide an estimate of a person's ability parameter by using any sample of items that measure the same trait, as opposed to classical test theory in which a person's true score is estimated by his or her observed score on the same set of test items presented to all of the examinees. These properties of the item response models, referred to as invariance by Lord (1980) and objectivity by Rasch (1966), mean that there exists what Wright (1968) refers to as sample-free item calibration and item-free person measurement.

d. Finally, item response models are appropriately applied to test data from power tests. Test data obtained under speeded conditions violate the assumption of unidimensionality since two traits influence test performance within a speeded test: speed and ability.

In addition to these four assumptions or properties there exist distinctions among the different item response mathematical models. As was stated previously, the item response models for dichotomously scored items differ according to the number of item parameters included in the model (the usual item parameters are those for difficulty, discrimination, and guessing), but there are certain characteristics that are similar in all of the models. Each model defines a mathematical function referred to as an item characteristic curve (ICC) that relates the probability of success on an item to the level of ability of the examinee. For dichotomously scored items, the number of item parameters (one, two, or three) needed for the definition of an ICC will depend upon the particular model.

The three-parameter logistic (3PL) model specifies three item parameters. One of the parameters, known as the discrimination level or 'a' parameter, is proportional to the

slope of the ICC at its steepest point or point of inflection (Hambleton & Swaminathan, 1985; Warm, 1978). This parameter typically ranges in value from 0 to +2 and is an index of how well the item discriminates between persons with different levels of ability. Another parameter of the 3PL model is the difficulty or 'b' parameter. It is defined as the point on the ability scale that corresponds to the point of inflection of the ICC. The item difficulty parameter is defined on the same ability (theta) scale as the person parameters and, in practice, typically ranges in value from -3 to +3. The more difficult the item, the greater the difficulty value. A third parameter is the pseudo-guessing index or 'c' parameter, which corresponds to the lower asymptote of the ICC. With some items, such as multiple-choice items, it is possible for a person with a very low ability level to answer a difficult item correctly purely by chance; this factor is taken into account by the 3PL model with the inclusion of the lower asymptote parameter.

The 2PL model is a special case of the 3PL model and assumes no guessing; the lower asymptote parameters for all ICCs are equal to zero. Therefore, only two item parameters, the a and b parameters, are included in the mathematical function for the 2PL model.

The 1PL or Rasch model is considered a special case of the 2PL and 3PL models. It assumes that all the items are equally discriminating among examinees, i.e., for all the items the slopes at the points of inflection are the same; and like the 2PL model, guessing does not exist. For the 1PL model, then, the only item parameter included in the mathematical function is the b parameter.

An important concept in item response theory contributed by Birnbaum (1968) is the information function. An information function reflects the accuracy of the ability estimates obtained from the item responses. The information function for a single item varies across the levels of ability. The higher the amplitude of the information curve, the greater the information. The contribution that a given item makes toward the effectiveness of measurement of the whole test is independent of what other items are included in the test (Birnbaum, 1968; Lord, 1980); therefore, item information is additive. The information of the total test is equal to the sum of the information functions for the items comprising the test. The use of item information functions in test construction is easy to illustrate. For example, if a test is to be administered to a population of people whose ability parameters lie within a particular range of values, a target information function (TIF) can be specified so precalibrated items can be selected to maximize the information in that particular ability range.

IRT models appear to have resolved some of the classical test model's shortcomings such as sample dependency; however,

problems in automated test construction do exist within these mathematical models. For example, computational procedures used in estimating the a, b, and c parameters are much more complicated than the estimation procedures of classical test theory indices, such as p-values, biserials, and reliability coefficients (Allen & Yen, 1979). The computer programs available for estimating item parameters and person parameters (thetas), through maximum likelihood or Bayesian routines, require a considerable amount of computer time as well as large numbers of examinees. Some programs require at least 60 test items and 1,000 subjects in order to produce reliable estimates.

Another shortcoming of IRT models is the indeterminacy of parameters. The invariance of the item parameter and ability parameter estimates holds only as long as the origin and unit of measurement of either the ability scale or the difficulty scale are fixed. Thus, there is an indeterminacy in the models in that the origin and unit of measurement used in any particular calibration are chosen arbitrarily. However, it is common practice or convention to choose the origin and unit of measurement for ability such that the estimated mean of the ability parameter estimates is zero and the estimated variance of the ability parameter estimates is one. Establishing the ability scale fixes simultaneously the unit of measurement of the item difficulty scale. Subsequent item parameter estimates will be invariant within a linear transformation. For example, if a common set of items is calibrated separately for two samples of people, the difficulty parameter estimates will not be identical; they will differ by a constant amount within a certain amount of error in estimation. This occurs because the difficulty scales have different origins but have the same unit of measurement. The discrimination parameter estimates, having a common origin, will be identical from group to group again within a certain amount of error in estimation, except for a change in the unit of measurement (Lord, 1975). The lower asymptote or c parameters are not affected by changes in the origin and unit of measurement of the ability scale, and therefore, should be relatively identical (there is some error in estimation) from one group to another. Within a group of persons, if each one takes two different tests measuring the same trait, each individual's ability scores will not be the same on the two tests. Like the difficulties, the ability scores will differ by a constant amount with a certain amount of error in estimation, again because the ability scales have different origins but have the same unit of measurement. Due to the common unit of measurement, it is easy to link these ability parameters onto a common scale.

## 3. Resolutions of Shortcomings in Automated Test Construction

As previously discussed, one of the major shortcomings of classical test models is sample dependency in the calculation of p-values, biserials and test reliability coefficients. Test development programs have addressed the sample dependency problem

by administering anchor test items along with experimental test items in a counter-balanced design. In the ASVAB development process this method is used, but only during the try-out phases of experimental items. In the initial administration of new items, anchor items, in this case, the reference test items from ASVAB Form 8a, are administered along with the experimental test items in a counter-balanced design. The classical statistics (p-values and biserials) for the experimental ASVAB items are matched with the corresponding 8a item statistics. Those items not matching the anchor test item statistics, as well as taxonomical categories, are deleted. After items have gone through the various try-out phases, operational length tests can be constructed that would be equivalent or parallel to each other as well as to the reference test. Subsequent administrations of operational length ASVAB forms, however, are not administered to the same sample. An equivalent groups design is used for these administrations. Therefore, the classical item statistics from these administrations are not comparable. Further, older operational length forms are not linked to newer operational length forms, again due to the use of an equivalent groups design. Therefore, for the current ASVAB development program, sample dependency in regard to the use of classical item statistics remains a problem.

With regard to item response models, the indeterminacy problem needs to be resolved in order to develop alternate forms using IRT item parameters. A resolution of this metric issue is a linear transformation of the a's and b's for one test to the scale of another test (recall that the c's are already on the same scale). Warm (1978) offers the formulae for these transformations using the means and standard deviations for each group of the a and b parameters.

The scaling and 'fit of the model' are other issues within IRT that can be handled in a manner similar to the item selection process for current production ASVAB forms. The first step would be to administer the new test items to a new group of examinees (a minimum number of examinees would be 500). Then the theta scale for the new sample is linearly transformed so that the items in the reference test recapture as closely as possible the a's and b's yielded by the new group of examinees. Once the theta scale for the new sample is calculated, the a's and b's can be estimated for the experimental test with the new sample. Both forms of the test would yield approximately the same parameter values. However, the standard errors of measurement may not be equal. To address the constraint of both forms producing the same standard errors at every value of theta, the new item parameters must again be matched with the parameters of the reference form. This method, as in classical test construction, is used to equate test forms (Allen & Yen, 1979).

Another issue that needs to be addressed is differential item functioning (item bias). Classical test construction does

not offer a good solution to differential item functioning (DIF) analyses. Difficulty values can be calculated for each subgroup of interest and compared, but these calculations, again, are dependent on the sample of examinees. However, IRT models are very useful in detecting ethnic or gender DIF. This is accomplished by estimating the item parameters separately for each subgroup. An item would be identified as functioning differently if for the same level of ability the probability of getting the item correct was different for each group. Items not conforming can be deleted from the overlength tests. A word of caution must be interjected at this point. Frequently the size of the subgroup is too small to offer any meaningful information about DIF, especially when using item response parameters. Therefore, it is anticipated that it may be valuable to incorporate classical DIF detection procedures, such as the Mantel-Haenszel procedure (Linn, Hastings, Hu, & Ryan, 1988).

In summary, the test developer, whether constructing test forms using classical or item response models, would need an automated item bank. The bank would need to contain not only classical and IRT statistics but also sample information (size, population descriptors, subgroup information, etc.) and test form identifiers (form number, administration site, type of administration instructions used, etc.). Each item would be identified and linked to its statistics, sample demographic information, and form characteristic information. As discussed, sample demographics and form characteristics should not be viewed merely as identifiers or links but should be noted and considered within the test development process. Once all relevant information is at hand, the on-line test constructor would be interested in the mathematical models that can accomplish the preparation of test forms and report meaningful parameters and scores within an automated system.

## B. Mathematical Modeling in Test Construction

Recent studies in the area of psychometrics have shown that the design of tests can be viewed as a 'decision process under certainty' which can be modeled with the techniques of mathematical programming (Van der Linden, 1987). Van der Linden presents the findings from three papers dealing with the mathematical models applied to the various aspects of test construction. The first paper, by Theunissen, addresses the test construction decision process in the simultaneous development of parallel forms. This process consists of options and conditions, where options imply the selection of some specified number of items and conditions consist of linear equalities and inequalities constraining the exercise of these options. Assuming there exists a pool of test items, finding a test consisting of items from this pool involves formulating specific constraints which limit the selection of items. Having selected a subset of these items which are admissible under the given constraints, the task remains to find the desired elements of this subset which

together form an optimal test. Theunissen uses the example of a hiker choosing objects to place inside a knapsack as a demonstration for the 'packing' class of problems. This example supposes that each object has a certain monetary value as well as a certain weight. In choosing which objects to pack and which to leave behind, the 'goal' or 'objective' is to minimize the total weight, subject to the constraint that the resulting total value of the objects selected must equal or exceed a certain dollar amount. Theunissen shows, briefly, how this simple concept of a weighted sum function can be extended to the problem of constructing a test with specified qualities.

This process resembles the present test development procedures of the enhanced AIB system (discussion of this system will follow). Items are viewed manually by the test developer, who must then implement constraints such as difficulty and discrimination value ranges, desired taxonomy categories, or the absence of negative biserial correlations for item distractors. These procedures are accomplished on paper or interactively; however, they could be accomplished just as well in an automated mode. The weights assigned to these constraints would be applied as the linear program scans through the item bank. However, the process would not be an interactive one and would result in the construction of only one test form. Thus, the issue of simultaneous construction of parallel forms needs to be considered.

The second paper by Boekkooi-Timminga (Van der Linden, 1987) gives an overview of simultaneous test construction methods using a special case of mathematical programming called zero-one programming. Using an item selection process based on the concept of information from IRT, Boekkooi-Timminga presented some objective functions and practical constraints which extend the ideas presented by Theunissen. Boekkooi-Timminga outlined three methods for test construction, using as constraints the same target test information function (TIF), no overlapping test items, and the same number of items for each test form. Of particular interest is the first method which assigns items to each test and measures the maximum difference between the tests' TIFs.

The most pertinent of the three methods presented by Boekkooi-Timminga seems to minimize the maximum distance between two constructed test versions and can be applied to the construction of new forms of the ASVAB. The on-line ASVAB test developer would be able to indicate multiple constraints (i.e., target TIFs, number of items per test, taxonomic classification, difficulty levels, discrimination levels, etc.). Items within the bank would be assigned to one test only and multiple test versions would be constructed. After the actual TIF values have been calculated for each new test and reference Form 8a, the program would determine the maximum distance between the TIFs for each test version. The desired result would be a minimum

distance among the new forms and a minimum distance between all new forms and Form 8a. However, as Boekkooi-Timminga stated, this procedure is computationally complex even with the 5-item test case. More research using longer tests and larger sample sizes is needed in order to test the algorithms and information estimations.

The third paper of interest, by Kelderman (Van der Linden, 1987), addressed the mathematical models that could be applied for remedying previously mentioned scaling problems. As with the other mathematical models, a target TIF must be specified. However, a satisfactory method for specifying a target TIF is not available. Thus, Kelderman presented other ways to interpret test information. First he explained that ability or theta can be related to quantities that are familiar to the test developer and the test user, such as percentiles for a reference population. The algorithms express the percentile point for a certain ability level in terms of the cumulative density function of theta in the population of interest (e.g., for ASVAB the population of interest is accessions).

These percentiles are then used in a paired-comparison method as a way of interpreting information. This method yields values of an information function for different scale points through wrong-order probabilities; i.e., individuals being falsely estimated as more "able" than other individuals who seem to possess more ability. In an interactive mode, the on-line test constructor would be presented with scale points, along a line, that are anchored to percentile points. The next step would involve the test developer's selecting an interval around the scale point of interest. The items corresponding to the endpoints of the interval would be highlighted. The test constructor would then indicate the wrong order probability level that could be tolerated for each of the endpoints of the interval. Kelderman presented a paired-comparison process that involves reversing this relationship and then calculating a new point on the scale from the old interval. This procedure is repeated for each scale point of interest until all relevant scale points are processed and corresponding items are highlighted. The new scale points would be used in constructing tests.

The shortcoming of this paired-comparison method lies in the use of the 1PL or Rasch model. The test developer would be interested in expanding the model to the 3PL case to include discrimination and pseudo-guessing parameters and would be interested in all scale points; however, such an expansion would prove computer-intensive. In addition, statistical checks would need to be built into the package to ensure the reliability of the subjective judgments of the developer.

In summary, the three papers presented here describe ways in which mathematical modeling can be used to simultaneously

construct parallel forms of a test. For most testing programs, developing tests on-line is preferable. On-line test construction is usually taken to mean an interactive process whereby the test constructor retrieves items from a pool by specifying certain parameter range restrictions and then reviews each item to decide whether or not to include the item in the test. The linear program packages presented by the authors in Van der Linden (1987) can be applied in order to construct parallel forms simultaneously; however, feasibility studies would need to be conducted in order to ascertain the toll on computer time and the reliability and validity of tests constructed using IRT.

## C. Prior Projects at AFHRL

Projects to develop an automated item banking and test construction method have been undertaken on several occasions at AFHRL throughout the past 10 years. Before emphasizing the current concerns and applications, it is appropriate to review two of the major accomplishments of the recent past.

### 1. 1978 AIB System

The 1978 version of the AIB system (Ree, 1978), written for the SPERRY-UNIVAC 1108 using ASCII-FORTRAN, consisted of three menu-driven program clusters which together provided a four-part banking procedure (a three-part test construction procedure and a one-part item pool editing procedure). Part 1, the test construction program, permitted two types of searches based on IRT or classical item parameters. Items were selected from the item pool, one at a time, and presented for review at the user terminal. A decision was made to include or not to include the item in the test being constructed. Another item was retrieved, and the process continued until the desired number of test items had been selected. Part 2 of this system provided a method to update the item pool such as recording that a particular item had been used in the operational test being constructed. Part 3 of the AIB system (the final feature of the test construction program) allowed the user to choose from among three output options or to save intermediate results for further work on the same test. Finally, Part 4 of the AIB system provided a second program much like an 'editor,' which allowed extensive revision of the individual items within the item pool.

In this 1978 version of the AIB system, the computation of test statistics was accomplished after all items for a particular test had been selected. Furthermore, this version made no comparison of the candidate items with the respective reference form items. Hence, if items of lower difficulty than those of the reference form were picked consistently throughout the test construction process, then only after test statistics were computed could the test constructor make appropriate compensatory adjustments.

## 2. 1983 AIB Enhancement

An enhancement of the AIB system by Lee and Fairbank (1983) addressed four areas. First, the item pool editing program was modified to include a batch mode for appending items to the item pool for use when large groups of items were to be added. The 1978 version was cumbersome in this respect. Second, a menu of available search options was added to allow for a wide variety of item parameters on which to base retrieval from the item pool. There was also a two-level search feature added which would permit a search on p-value, biserial, keyword, or a combination of any two of these. Lee and Fairbank's intent was to allow for the easy addition of other parameter searches. The items could be recovered as a group, with each retrieved item having the specified parameter within the designated range. These retrieved items could be displayed to the user in list format, along with important item characteristics, thus allowing flexibility in the selection of items for a specific form of the test. Also when selecting an item for inclusion in the test, it is helpful to know the effect the item will have on the characteristics of the total test. A graphic display of both the Item Information Curve and a Test Information Curve was added as an aid to the test constructor. Finally, the desire for portability required a program written in a language which was supported by a wide range of computer systems. To this end, ASCII-FORTRAN was used to code all modules of the enhanced AIB system, thus eliminating the calls to UNIVAC-Assembler routines found in the 1978 version.

# III. CURRENT CONCERNS

The 1983 enhancement was successful to the extent that it provided both item banking and test development capabilities. Although the conversion to ASCII-FORTRAN made the AIB package much more portable, the absence of UNIVAC-specific routines found in the 1978 Fieldata version made the 1983 version somewhat slower and less efficient. The graphics routines required for the display of Item and Test Information Curves were unique to the Tektronix terminal equipment in use at that time and made the package less than fully portable. Hence, the search continues for a means of automated item banking and test construction which is both versatile and efficient while also being fully portable and compatible with different types and sizes of equipment.

Since ASVAB test development has been performed over the years by numerous contractors and the items used in test construction have been developed using standard procedures, there has been no concentrated effort to gather candidate items into a pool for common use. Furthermore, recent policy changes in the area of military test development will allow the reuse of a certain percentage of previously used items in the construction

of new subtests. Another concern of the present tasking, therefore, is to provide a database of items, including those previously used on all existing operational and nonoperational forms of the ASVAB and those items being written for the development of the next generation of ASVAB versions (Forms 20, 21, and 22). For all entries, item and test statistics must be included, where available, so that the resulting pool can be truly useful in the automated construction of parallel forms.

In light of recent developments in the area of mathematical programming (MP), another possibility at this time involves the prospect of using the MP approach with on-line interaction by the test constructor. The key to a successful MP approach is to provide the capability to incorporate all the varied concerns of the test constructor into solvable linear objectives and constraints, and to apply this model within a software and hardware system that will simultaneously develop parallel tests forms in a reasonable period of computer time.

An additional concern of the current effort is the simultaneous construction of more than one test form, with each form being parallel to a reference form. Although earlier efforts have accomplished this goal through repeated development of single forms, a means of building all forms simultaneously, and so making use of all items, remains a challenge; however, the MP approach is a viable solution to this concern.

Added to all of the concerns previously discussed is the requirement--perhaps unique to ASVAB developers--that the resultant parallel forms be printed with the same type font, pitch, spacing, and format as those of the reference test. To address these concerns, the item bank and test development program must interface with a publishing system that will generate paper copies of tests with these specifications as well as the associated item illustrations and mechanicals.

In view of the scope of this particular effort, the item banking and test construction problem is currently viewed as requiring a three-phase approach. Phase I is the development of a computerized item bank that allows items meeting specified criteria to be selected by the on-line test developer. Phase II is the expansion of this concept to include the automated simultaneous construction of parallel forms. Finally, Phase III is the further enhancement of the resulting system to include publishing of the actual test booklets from information stored in the computer. With this three-phase approach in mind, the remainder of this paper focuses on the specific content requirements of a comprehensive item banking system which will facilitate the natural progression to Phases II and III. It further considers the specification of the hardware and software currently deemed most appropriate to the efforts of Phase I and most suitable to such future expansion.

## IV. CONTENT REQUIREMENTS

The information which is fundamental to this effort can be considered to be of two types:

1. characteristics of the items themselves, and
2. statistics pertaining either to an experimental or to an operational use of an item.

Note that the content requirements for item bank files have used the ASVAB and the Air Force Officer Qualifying Test (AFOQT) as examples; however, these requirements pertain to other test development programs as well.

### A. Item Characteristics

Characteristics of the item include:

1. text of the item stem, correct response, and distractors;
2. for reading comprehension items (such as Paragraph Comprehension within the ASVAB), the text of the paragraph to which the item belongs, as well as printing dimensions;
3. position of the correct response (i.e., a, or b, or c, or d, etc.);
4. taxonomic classification;
5. whether or not the item requires an illustration and, if so, a reference to an image of this drawing, including printing dimensions;
6. words or phrases which relate the item to other items in the database (key words);
7. flags to other items which should not be used with this item on the same subtest form (mutually exclusive);
8. whether or not the item has been reviewed for ethnic/gender sensitivity;
9. original author (individual or firm) of the item;
10. whether the item has been used operationally, if so, test number and date of last operational use;
11. number of times the item has been operationally used;
12. comments regarding any modifications historically associated with this item.

### B. Item Statistics

Statistics pertaining to either an experimental or an operational use of the item encompass:

1. type of subject sample (Air Force recruit training center, AF-RTC; all Service RTC; Military Entrance Processing Station, MEPS; High School, HS; 1980

National Opinion Research Center, NORC; Officer Training School, OTS; Air Force Reserve Officer Training Corps; etc.);

2. location (site) of subjects in the sample;
3. size of the total sample;
4. development phase of the item (experimental tryout, ET; overlength; operational calibration, OPCAL; initial operational test and evaluation, IOT&E; in operation; reference);
5. date of testing;
6. indication of the set of directions used when administering this item;
7. name or identifier used on the experimental or operational form in which this item appears;
8. indication of whether the item was used in a speeded, power, or mixed subtest;
9. indication of whether the item was used as an anchor item and, if so, whether or not there exists an appropriate link to the corresponding statistics;
10. indication of whether an item was 'scored' for use in the computation of an individual's composite(s);
11. length of the form (overlength, production length, etc.);
12. relative number of the item within this form;
13. position of the correct alternative (key);
14. position of each of the distractors;
15. p-value and R-biserial of the correct alternative with the associated standard error of measurement;
16. correlation coefficient corrections (restriction of range, etc.);
17. selected test statistics (mean, variance, skew, kurtosis, etc.);
18. classical item statistics for the total sample (R-biserial and p-value), presented for each response alternative;
19. IRT parameters for the total sample with the associated standard error of measurement for each parameter;
20. classical item statistics for the population subgroups (male, female, white, Black, and Hispanic), presented for each response alternative;
21. IRT parameters for the population subgroups; and
22. comments regarding any modifications historically associated with this use of the item.

## V. STRUCTURE OF THE DATABASE SYSTEM

Regardless of the hardware and software chosen to serve as host, the item bank should be structured as a relational database system, with each independent database corresponding to a

specific subtest. Within each subtest database, a file of items should contain fields for each of the item characteristics discussed above. Additionally, a set of linked files (records are linked one-to-one) should contain fields for each of the item statistics also addressed above. Figure 1 shows the proposed positional relationship of the files and databases within this system. The linked statistics files would contain one record for each instance in which an item was used either experimentally or operationally. Hence, one or more records might relate to a single record within the item characteristic file. The database is thus relational in that a single record in one file is related (one-to-many) to one or more records in a sibling file or files.

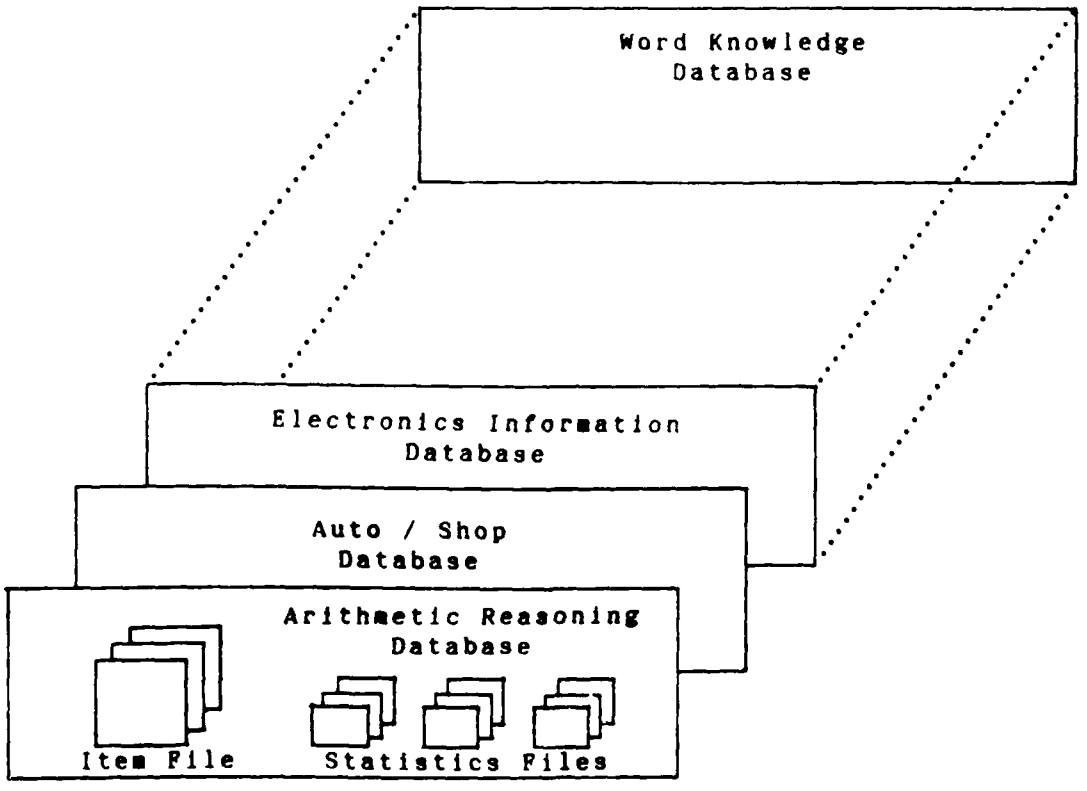



Figure 1. Item Bank Database System.

Figures 2 through 5 show the recommended specific layout of the Item Characteristics File (Figure 2) and each of the three Item Statistics Files (Figures 3, 4 and 5) using the Arithmetic Reasoning (AR) subtest as an example.

| Field | Field Name | Type | Width | Dec | Explanation |
|---|---|---|---|---|---|
| 1 | ITEM_ID | C[a] | 5 | | Unique item identification. |
| 2 | STEM | C | 250 | | Text for item stem. |
| 3 | CORRECT | C | 60 | | Text for correct response. |
| 4 | DISTR01 | C | 60 | | Text for distractor 1 of 3. |
| 5 | DISTR02 | C | 60 | | Text for distractor 2 of 3. |
| 6 | DISTR03 | C | 60 | | Text for distractor 3 of 3. |
| 7 | TAXONOMY | C | 2 | | Item's taxonomic category. |
| 8 | KEYWORD01 | C | 60 | | Text for keyword/phrase 1 of 3. |
| 9 | KEYWORD02 | C | 60 | | Text for keyword/phrase 2 of 3. |
| 10 | KEYWORD03 | C | 60 | | Text for keyword/phrase 3 of 3. |
| 11 | METAG01 | C | 5 | | Item shouldn't be used with ... |
| 12 | METAG02 | C | 5 | | Item shouldn't be used with ... |
| 13 | METAG03 | C | 5 | | Item shouldn't be used with ... |
| 14 | AUTHOR | C | 60 | | Name of item's original author. |
| 15 | TIMES_USED | C | 2 | | Number of times item was used. |
| 16 | LAST_USED | D | 8 | | Date of item's last use. |
| 17 | REMARKS | M | 10 | | Additional remarks or comments. |
| Total Character Width | | | 773 | | |

[a] C = Character
D = Date
M = Memo

Figure 2. Example of Item Characteristics File: ASVAB Arithmetic Reasoning Items.

Other subtest examples, such as Mechanical Comprehension (MC), would require a field in the Item Characteristics File to indicate illustrations. Items requiring illustrations or mechanical symbols would contain either a reference in the Item Characteristics File that points to a printed sample of the appropriate drawing or, in a more sophisticated publishing environment, the actual graphic representation of this depiction. In terms of the three-phase approach of the item banking and test construction effort mentioned earlier, early implementations of the item bank could contain a simple reference to the graphics while future elaborations could include the storing of the actual mechanicals (rub-on letters/numbers) and illustrations so that it would be possible to publish the tests directly from the item bank.

| Field | Field Name | Type | Width | Dec | Explanation |
|---|---|---|---|---|---|
| 1 | ITEM_ID | C[a] | 5 | | Unique item identification. |
| 2 | STAT_LINK | C | 5 | | Link to parts 2,3 of item stats. |
| 3 | SAMPL_SITE | C | 60 | | Location of subjects in sample. |
| 4 | PHASE | C | 10 | | Development phase of this test. |
| 5 | TESTDATE | D | 8 | | Date of this sample testing. |
| 6 | TESTNAME | C | 6 | | Name used to identify test. |
| 7 | TESTLENGTH | C | 2 | | Length of form for this admin. |
| 8 | ITEM_NUMBR | C | 2 | | Item's number on this admin. |
| 9 | CORRECT_AT | C | 1 | | Position of correct alter'tive. |
| 10 | DISTR01_AT | C | 1 | | Position of distractor 1. |
| 11 | DISTR02_AT | C | 1 | | Position of distractor 2. |
| 12 | DISTR03_AT | C | 1 | | Position of distractor 3. |
| 13 | CORRECT_P | N | 6 | 3 | P-value of correct alternative. |
| 14 | CORRECT_B | N | 6 | 3 | R-biserial of correct alt'tive. |
| 15 | SAMPL_SIZE | C | 5 | | Size of total sample. |
| 16 | TEST_MEAN | N | 7 | 3 | Test mean raw score. |
| 17 | TEST_VARCE | N | 7 | 3 | Test raw score variance. |
| 18 | TEST_SKEW | N | 7 | 3 | Test raw score measure of skew. |
| 19 | TEST_KURT | N | 7 | 3 | Test raw score kurtosis. |
| 20 | RAW_MINMUM | N | 7 | 3 | Observed raw score minimum. |
| 21 | RAW_MAXMUM | N | 7 | 3 | Observed raw score maximum. |
| 22 | RAW_MEDIAN | N | 7 | 3 | Observed raw score median. |
| 23 | TEST_STDEV | N | 7 | 3 | Test raw score std. deviation. |
| 24 | TEST_KR20 | N | 7 | 3 | Test alpha reliability coeff't. |
| 25 | TEST_SEM | N | 7 | 3 | Test standard error of meas't. |
| 26 | MEAN_PVAL | N | 7 | 3 | Test average P-value. |
| 27 | MEAN_RBIS | N | 7 | 3 | Test average R-biserial. |
| 28 | MEAN_PTBIS | N | 7 | 3 | Test average point biserial. |
| 29 | REMARKS | M | 10 | | Additional remarks or comments. |
| Total Character Width | | | 221 | | |

[a] C = Character
D = Date
N = Numeric
M = Memo

Figure 3. Example of First Item Statistics File: ASVAB Arithmetic Reasoning Items.

## VI. HARDWARE AND SOFTWARE ALTERNATIVES

A survey of the present technologies reveals several hardware and software systems which could serve as host to the item bank and would enable the test developer to automate test construction and booklet publication. Certainly, the possibilities to be considered include both a mainframe computer used in conjunction with the numerous item and test analysis software packages developed over the last several decades, and a microcomputer used in conjunction with one of the various

| Field | Field Name | Type | Width | Dec | Explanation |
|---|---|---|---|---|---|
| 1 | ITEM_ID | C[a] | 5 | | Unique item identification. |
| 2 | STAT_LINK | C | 5 | | Link to parts 1,3 of item stats. |
| 3 | T_RBIS_O | N | 6 | 3 | R-bis, total sample, omits. |
| 4 | T_PVAL_O | N | 6 | 3 | P-val, total sample, omits. |
| 5 | T_NNNN_O | N | 5 | | Number selecting omit (total). |
| 6 | T_MEAN_O | N | 6 | 3 | Mean Z-scr, omits (total). |
| 7 | T_RBIS_A | N | 6 | 3 | R-bis, total sample, alt've -A. |
| 8 | T_PVAL_A | N | 6 | 3 | P-val, total sample, alt've -A. |
| 9 | T_NNNN_A | N | 5 | | Number selecting -A (total). |
| 10 | T_MEAN_A | N | 6 | 3 | Mean Z-scr, A-resp'nts (total). |
| 11 | T_RBIS_B | N | 6 | 3 | R-bis, total sample, alt've -B. |
| 12 | T_PVAL_B | N | 6 | 3 | P-val, total sample, alt've -B. |
| 13 | T_NNNN_B | N | 5 | | Number selecting -B (total). |
| 14 | T_MEAN_B | N | 6 | 3 | Mean Z-scr, B-resp'nts (total). |
| 15 | T_RBIS_C | N | 6 | 3 | R-bis, total sample, alt've -C. |
| 16 | T_PVAL_C | N | 6 | 3 | P-val, total sample, alt've -C. |
| 17 | T_NNNN_C | N | 5 | | Number selecting -C (total). |
| 18 | T_MEAN_C | N | 6 | 3 | Mean Z-scr, C-resp'nts (total). |
| 19 | T_RBIS_D | N | 6 | 3 | R-bis, total sample, alt've -D. |
| 20 | T_PVAL_D | N | 6 | 3 | P-val, total sample, alt've -D. |
| 21 | T_NNNN_D | N | 5 | | Number selecting -D (total). |
| 22 | T_MEAN_D | N | 6 | 3 | Mean Z-scr, D-resp'nts (total). |
| 23 | T_PARAMA | N | 6 | 3 | IRT parameter-A, total sample. |
| 24 | T_PARAMB | N | 6 | 3 | IRT parameter-B, total sample. |
| 25 | T_PARAMC | N | 6 | 3 | IRT parameter-C, total sample. |
| 26 | M_RBIS_O | N | 6 | 3 | R-bis, male sample, omits. |
| 27 | M_PVAL_O | N | 6 | 3 | P-val, male sample, omits. |
| 28 | M_NNNN_O | N | 5 | | Number selecting omit (male). |
| 29 | M_MEAN_O | N | 6 | 3 | Mean Z-scr, omits (male). |
| 30 | M_RBIS_A | N | 6 | 3 | R-bis, male sample, alt've -A. |
| 31 | M_PVAL_A | N | 6 | 3 | P-val, male sample, alt've -A. |
| 32 | M_NNNN_A | N | 5 | | Number selecting -A (male). |
| 33 | M_MEAN_A | N | 6 | 3 | Mean Z-scr, A-resp'nts (male). |
| 34 | M_RBIS_B | N | 6 | 3 | R-bis, male sample, alt've -B. |
| 35 | M_PVAL_B | N | 6 | 3 | P-val, male sample, alt've -B. |
| 36 | M_NNNN_B | N | 5 | | Number selecting -B (male). |
| 37 | M_MEAN_B | N | 6 | 3 | Mean Z-scr, B-resp'nts (male). |
| 38 | M_RBIS_C | N | 6 | 3 | R-bis, male sample, alt've -C. |
| 39 | M_PVAL_C | N | 6 | 3 | P-val, male sample, alt've -C. |
| 40 | M_NNNN_C | N | 5 | | Number selecting -C (male). |

Figure 4. Example of Second Item Statistics File: ASVAB Arithmetic Reasoning Items.

| Field | Field Name | Type | Width | Dec | Explanation |
|---|---|---|---|---|---|
| 41 | M_MEAN_C | N[a] | 6 | 3 | Mean Z-scr, C-resp'nts (male). |
| 42 | M_RBIS_D | N | 6 | 3 | R-bis, male sample, alt've -D. |
| 43 | M_PVAL_D | N | 6 | 3 | P-val, male sample, alt've -D. |
| 44 | M_NNNN_D | N | 5 | | Number selecting -D (male). |
| 45 | M_MEAN_D | N | 6 | 3 | Mean Z-scr, D-resp'nts (male). |
| 46 | M_PARAMA | N | 6 | 3 | IRT parameter-A, male sample. |
| 47 | M_PARAMB | N | 6 | 3 | IRT parameter-B, male sample. |
| 48 | M_PARAMC | N | 6 | 3 | IRT parameter-C, male sample. |
| 49 | F_RBIS_O | N | 6 | 3 | R-bis, fmale sample, omits. |
| 50 | F_PVAL_O | N | 6 | 3 | P-val, fmale sample, omits. |
| 51 | F_NNNN_O | N | 5 | | Number selecting omit (fmale). |
| 52 | F_MEAN_O | N | 6 | 3 | Mean Z-scr, omits (fmale). |
| 53 | F_RBIS_A | N | 6 | 3 | R-bis, fmale sample, alt've -A. |
| 54 | F_PVAL_A | N | 6 | 3 | P-val, fmale sample, alt've -A. |
| 55 | F_NNNN_A | N | 5 | | Number selecting -A (fmale). |
| 56 | F_MEAN_A | N | 6 | 3 | Mean Z-scr, A-resp'nts (fmale). |
| 57 | F_RBIS_B | N | 6 | 3 | R-bis, fmale sample, alt've -B. |
| 58 | F_PVAL_B | N | 6 | 3 | P-val, fmale sample, alt've -B. |
| 59 | F_NNNN_B | N | 5 | | Number selecting -B (fmale). |
| 60 | F_MEAN_B | N | 6 | 3 | Mean Z-scr, B-resp'nts (fmale). |
| 61 | F_RBIS_C | N | 6 | 3 | R-bis, fmale sample, alt've -C. |
| 62 | F_PVAL_C | N | 6 | 3 | P-val, fmale sample, alt've -C. |
| 63 | F_NNNN_C | N | 5 | | Number selecting -C (fmale). |
| 64 | F_MEAN_C | N | 6 | 3 | Mean Z-scr, C-resp'nts (fmale). |
| 65 | F_RBIS_D | N | 6 | 3 | R-bis, fmale sample, alt've -D. |
| 66 | F_PVAL_D | N | 6 | 3 | P-val, fmale sample, alt've -D. |
| 67 | F_NNNN_D | N | 5 | | Number selecting -D (fmale). |
| 68 | F_MEAN_D | N | 6 | 3 | Mean Z-scr, D-resp'nts (fmale). |
| 69 | F_PARAMA | N | 6 | 3 | IRT parameter-A, fmale sample. |
| 70 | F_PARAMB | N | 6 | 3 | IRT parameter-B, fmale sample. |
| 71 | F_PARAMC | N | 6 | 3 | IRT parameter-C, fmale sample. |
| Total Character Width | | | 410 | | |

[a] C = Character
N = Numeric

Figure 4 (concluded)

available document publishing systems. For the sake of completeness, it is useful to examine some of the advantages and disadvantages of these environments more closely.

## A. Mainframe

The mainframe computer is superior to the microcomputer with respect to high-speed computation and mass data storage. Over the years, mainframe computers have become faster and more versatile. High-speed mass storage devices have evolved to the point where huge quantities of data can be stored and instantly

accessed for a multitude of analytic applications. The evolution of high-level languages, from the early days of FORTRAN, BASIC, and ALGOL to the currently popular ADA and PASCAL, continues to result in a wide range of tools for manipulating data and automating those processes once tediously done by hand.

| Field | Field Name | Type | Width | Dec | Explanation |
|---|---|---|---|---|---|
| 1 | ITEM_ID | C[a] | 5 | | Unique item identification. |
| 2 | STAT_LINK | C | 5 | | Link to parts 1,2 of item stats. |
| 3 | W_RBIS_O | N | 6 | 3 | R-bis, white sample, omits. |
| 4 | W_PVAL_O | N | 6 | 3 | P-val, white sample, omits. |
| 5 | W_NNNN_O | N | 5 | | Number selecting omit (white). |
| 6 | W_MEAN_O | N | 6 | 3 | Mean Z-scr, omits (white). |
| 7 | W_RBIS_A | N | 6 | 3 | R-bis, white sample, alt've -A. |
| 8 | W_PVAL_A | N | 6 | 3 | P-val, white sample, alt've -A. |
| 9 | W_NNNN_A | N | 5 | | Number selecting -A (white). |
| 10 | W_MEAN_A | N | 6 | 3 | Mean Z-scr, A-resp'nts (white). |
| 11 | W_RBIS_B | N | 6 | 3 | R-bis, white sample, alt've -B. |
| 12 | W_PVAL_B | N | 6 | 3 | P-val, white sample, alt've -B. |
| 13 | W_NNNN_B | N | 5 | | Number selecting -B (white). |
| 14 | W_MEAN_B | N | 6 | 3 | Mean Z-scr, B-resp'nts (white). |
| 15 | W_RBIS_C | N | 6 | 3 | R-bis, white sample, alt've -C. |
| 16 | W_PVAL_C | N | 6 | 3 | P-val, white sample, alt've -C. |
| 17 | W_NNNN_C | N | 5 | | Number selecting -C (white). |
| 18 | W_MEAN_C | N | 6 | 3 | Mean Z-scr, C-resp'nts (white). |
| 19 | W_RBIS_D | N | 6 | 3 | R-bis, white sample, alt've -D. |
| 20 | W_PVAL_D | N | 6 | 3 | P-val, white sample, alt've -D. |
| 21 | W_NNNN_D | N | 5 | | Number selecting -D (white). |
| 22 | W_MEAN_D | N | 6 | 3 | Mean Z-scr, D-resp'nts (white). |
| 23 | W_PARAMA | N | 6 | 3 | IRT parameter-A, white sample. |
| 24 | W_PARAMB | N | 6 | 3 | IRT parameter-B, white sample. |
| 25 | W_PARAMC | N | 6 | 3 | IRT parameter-C, white sample. |
| 26 | B_RBIS_O | N | 6 | 3 | R-bis, black sample, omits. |
| 27 | B_PVAL_O | N | 6 | 3 | P-val, black sample, omits. |
| 28 | B_NNNN_O | N | 5 | | Number selecting omit (black). |
| 29 | B_MEAN_O | N | 6 | 3 | Mean Z-scr, omits (black). |
| 30 | B_RBIS_A | N | 6 | 3 | R-bis, black sample, alt've -A. |
| 31 | B_PVAL_A | N | 6 | 3 | P-val, black sample, alt've -A. |
| 32 | B_NNNN_A | N | 5 | | Number selecting -A (black). |
| 33 | B_MEAN_A | N | 6 | 3 | Mean Z-scr, A-resp'nts (black). |
| 34 | B_RBIS_B | N | 6 | 3 | R-bis, black sample, alt've -B. |
| 35 | B_PVAL_B | N | 6 | 3 | P-val, black sample, alt've -B. |
| 36 | B_NNNN_B | N | 5 | | Number selecting -B (black). |
| 37 | B_MEAN_B | N | 6 | 3 | Mean Z-scr, B-resp'nts (black). |
| 38 | B_RBIS_C | N | 6 | 3 | R-bis, black sample, alt've -C. |
| 39 | B_PVAL_C | N | 6 | 3 | P-val, black sample, alt've -C. |
| 40 | B_NNNN_C | N | 5 | | Number selecting -C (black). |

Figure 5. Example of Third Item Statistics File: ASVAB Arithmetic Reasoning Items.

| Field | Field Name | Type | Width | Dec | Explanation |
|---|---|---|---|---|---|
| 41 | B_MEAN_C | N[a] | 6 | 3 | Mean Z-scr, C-resp'nts (black). |
| 42 | B_RBIS_D | N | 6 | 3 | R-bis, black sample, alt've -D. |
| 43 | B_PVAL_D | N | 6 | 3 | P-val, black sample, alt've -D. |
| 44 | B_NNNN_D | N | 5 | | Number selecting -D (black). |
| 45 | B_MEAN_D | N | 6 | 3 | Mean Z-scr, D-resp'nts (black). |
| 46 | B_PARAMA | N | 6 | 3 | IRT parameter-A, black sample. |
| 47 | B_PARAMB | N | 6 | 3 | IRT parameter-B, black sample. |
| 48 | B_PARAMC | N | 6 | 3 | IRT parameter-C, black sample. |
| 49 | H_RBIS_O | N | 6 | 3 | R-bis, hisp. sample, omits. |
| 50 | H_PVAL_O | N | 6 | 3 | P-val, hisp. sample, omits. |
| 51 | H_NNNN_O | N | 5 | | Number selecting omit (hisp.). |
| 52 | H_MEAN_O | N | 6 | 3 | Mean Z-scr, omits (hisp.). |
| 53 | H_RBIS_A | N | 6 | 3 | R-bis, hisp. sample, alt've -A. |
| 54 | H_PVAL_A | N | 6 | 3 | P-val, hisp. sample, alt've -A. |
| 55 | H_NNNN_A | N | 5 | | Number selecting -A (hisp.). |
| 56 | H_MEAN_A | N | 6 | 3 | Mean Z-scr, A-resp'nts (hisp.). |
| 57 | H_RBIS_B | N | 6 | 3 | R-bis, hisp. sample, alt've -B. |
| 58 | H_PVAL_B | N | 6 | 3 | P-val, hisp. sample, alt've -B. |
| 59 | H_NNNN_B | N | 5 | | Number selecting -B (hisp.). |
| 60 | H_MEAN_B | N | 6 | 3 | Mean Z-scr, B-resp'nts (hisp.). |
| 61 | H_RBIS_C | N | 6 | 3 | R-bis, hisp. sample, alt've -C. |
| 62 | H_PVAL_C | N | 6 | 3 | P-val, hisp. sample, alt've -C. |
| 63 | H_NNNN_C | N | 5 | | Number selecting -C (hisp.). |
| 64 | H_MEAN_C | N | 6 | 3 | Mean Z-scr, C-resp'nts (hisp.). |
| 65 | H_RBIS_D | N | 6 | 3 | R-bis, hisp. sample, alt've -D. |
| 66 | H_PVAL_D | N | | 3 | P-val, hisp. sample, alt've -D. |
| 67 | H_NNNN_D | N | 5 | | Number selecting -D (hisp.). |
| 68 | H_MEAN_D | N | 6 | 3 | Mean Z-scr, D-resp'nts (hisp.). |
| 69 | H_PARAMA | N | 6 | 3 | IRT parameter-A, hisp. sample. |
| 70 | H_PARAMB | N | 6 | 3 | IRT parameter-B, hisp. sample. |
| 71 | H_PARAMC | N | 6 | 3 | IRT parameter-C, hisp. sample. |
| Total Character Width | | | 410 | | |

[a] C = Character
N = Numeric

Figure 5 (Concluded)

With respect to automated item banking and test construction, a predominance of the literature and software published in the last decade shows that the mainframe computer continues to be the mainstay of the psychometric community. From simple item analysis, through development of simultaneously parallel forms via linear programming, to final calibration and equating of test results, the mainframe environment offers a wealth of programming and analytic devices.

Just as the advantages of such a powerful device are numerous and convincing, so too are the drawbacks of dependence

on such a machine. Because of the expense of maintaining the typical mainframe, a single machine is relied upon by many users and "downtime" is inevitable. At times when the mainframe is inaccessible, it is unlikely that an identical backup system will be handy and most software, once implemented on a particular mainframe, is not readily transportable to another. This can certainly affect the productivity of the test developer and those responsible for publishing the test booklets.

## B. Document Publishing System

Document publishing systems are becoming widespread throughout Government and Industry, with the publishing continuum ranging from a simple word processing workstation to a host-based document production environment. The cut-and-paste activities of traditional document production have been made obsolete by these emerging systems. The growing base of publishing hardware and software, though often incompatible among vendors, makes the business of desktop publishing an attractive option for the automated test construction process.

At the time of this report, many hardware and software vendors are responding to the need for an integrated document processing environment by providing industry-standard networking and communications links for an entire line of publishing products. Systems have already been demonstrated that can provide text-and-graphics output on a whole family of laser printers, all of which yield a high quality of print resolution. Sophisticated publishing software provides powerful page layout and document composition capabilities. These same PC-based publishing packages make it possible to combine text, data, and illustrations into a single document from word processors, spreadsheets, graphics packages and many other applications. Moreover, some of these same systems support a wide array of PC peripherals such as scanners, color monitors, color printers, plotters, and photo-typesetters.

One such system is the Signature Electronic Publishing System designed by VariTyper. A configuration which would be pertinent for developing future forms of the ASVAB, AFOQT, or other military test might consist of PC workstations for item entry and editing using a database management package such as dBaseIII Plus; a large fixed disk for item and illustration storage; an image scanner for both text and graphic entry; a laser printer for high-quality booklet printing; and the software required for networking, document editing and assembly, and publishing of the final test booklet. With a controlling system that is essentially a 286 or 386 microcomputer, this same PC-based environment might be used for the analytic portion of test development including item analysis, automated test construction, and test calibration and equating by simply adding the appropriate software to accomplish these tasks (e.g., the 1983 version of AIB). A 286 or 386 microcomputer provides the memory

necessary to perform time-consuming calculations vice a personal computer bought solely for word processing capabilities.

## VII. RECOMMENDATIONS

The files of an automated item banking system can be constructed using one of several packages created for use on the microcomputer. A widely used and well-respected package is the dBASE (tm) database software developed in 1985 by the Ashton-Tate software development firm. Item characteristic and statistic records may be manipulated (appended, edited, displayed, browsed, replaced, deleted, located, retrieved, sorted, etc.) using the well-documented features of this package which are easily used by someone not familiar with the dBASE language. To illustrate the contents of these files, Figures 6 and 7 have been included to show how dBase might display selected fields and records within both the Item Characteristics File (Figure 6) and the Item Statistics Files (Figure 7) for AR items. The information presented in these examples is for illustrative purposes and does not represent exhaustive lists of the information contained nor a comprehensive repertoire of the reports available from this database system.

Although much of the preliminary building of a comprehensive item bank is feasible and practical using a PC and the database software described above, it appears that the best and most cost-effective solution for Phases II and III will be a full-featured professional electronic publishing system consisting of PC workstations, a 286-based network server, and the other peripherals previously mentioned. A 286-based network server would store all the files; the peripherals would be able to access the files and perform the test development and publishing functions of the system. The capabilities of such a system would include provisions for electronically scanned entry of text and graphics and automatic spelling checks and correction of text, resulting in a comprehensive item banking system with all of the necessary demographic and statistical information necessary for fully automated test construction. The abundance of microcomputer software currently available makes the Phase II goal of parallel form construction on the same 286-based machine quite practical and feasible. The high-quality printing of both text and graphics, automatically re-sized to the correct dimensions and printed using the same type, font, pitch, and spacing of the reference test form, is a natural outcome of this configuration and a fitting solution to the objectives of Phase III.

Although most of the analyses required for automated construction of parallel forms are already possible using the features of the microcomputer, it is certainly conceivable that it may be either necessary or desirable to have some portion of the item bank database information available to the mainframe environment. The selection of microcomputer database software

| Item ID | Item Stem | Correct Choice | Distractors 1 | 2 | 3 |
|---|---|---|---|---|---|
| 0001 | Jim is 25 years old. Sarah is 8 years old. How many years older than Sarah is Jim? | 17 | 15 | 16 | 18 |
| 0002 | Copper tubing sells for 30 cents per yard. How many yards can be bought for $9.60? | 32 | 20 | 30 | 24 |
| 0003 | If apples cost 23 cents a pound, how many pounds can be bought for 69 cents? | 3 | 2 | 4 | 5 |
| 0004 | An airplane travels 63 miles in 20 minutes. What is the average speed of the plane, in miles per hour? | 189 | 162 | 211 | 224 |
| 0005 | What is a student's average in science if the student receives the following grades on tests: 93, 97, 84, 96, 78, and 80? | 88 | 90 | 92 | 96 |
| 0006 | A factory now employs 18 people. This is 60% fewer people than it previously employed. How many people did it previously employ? | 30 | 24 | 26 | 28 |
| 0007 | A bucket is filled with 8 gallons of a liquid that weighs 61 pounds. The bucket when it is empty weighs 5 pounds. How many pounds does 1 gallon of the liquid weigh? | 7 | 4 | 6 | 9 |
| 0008 | The width of a rectangle is 4 inches and the perimeter is 24 inches. What is the length of the rectangle? | 8 | 4 | 7 | 12 |
| 0009 | Postage on parcel A is $1.50. Postage on parcel B is 2/5 of the postage on A. What is the postage on parcel B in cents? | 60 | 30 | 40 | 50 |
| 0010 | A family took a 800-mile trip. First they traveled a certain number of miles by car, and then 7 times as far by airplane. How many miles did they travel by airplane? | 100 | 150 | 180 | 200 |

Figure 6. Sample Content of ASVAB Arithmetic Reasoning Items in the Item Characteristics File.

| Item ID | Testing Location | Devel Phase | Test Date | Name | Leng | No | Key | 1 | 2 | 3 | Key P-val | Key R-bis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0001 | MEPS | IOT/E | 10/79 | 25A | 30 | 01 | C | A | B | D | 0.710 | 0.599 |
| 0002 | MEPS | IOT/E | 10/79 | 25A | 30 | 02 | C | A | B | D | 0.965 | 0.241 |
| 0003 | MEPS | IOT/E | 10/79 | 25A | 30 | 03 | B | A | C | D | 0.634 | 0.345 |
| 0004 | MEPS | IOT/E | 10/79 | 25A | 30 | 04 | B | A | C | D | 0.803 | 0.462 |
| 0005 | MEPS | IOT/E | 10/79 | 25A | 30 | 05 | A | B | C | D | 0.860 | 0.338 |
| 0006 | MEPS | IOT/E | 10/79 | 25A | 30 | 06 | D | A | B | C | 0.775 | 0.179 |
| 0007 | MEPS | IOT/E | 10/79 | 25A | 30 | 07 | C | A | B | D | 0.742 | 0.245 |
| 0008 | MEPS | IOT/E | 10/79 | 25A | 30 | 08 | C | A | B | D | 0.733 | 0.237 |
| 0009 | MEPS | IOT/E | 10/79 | 25A | 30 | 09 | D | A | B | C | 0.581 | 0.306 |
| 0010 | MEPS | IOT/E | 10/79 | 25A | 30 | 10 | A | B | C | D | 0.691 | 0.372 |

Figure 7. Sample Statistics for ASVAB Arithmetic Reasoning Items.

places no limit on the potential sharing of these item data. Most major software vendors provide a means of exporting information from internal data files to standard text files (ASCII). ASCII files can then be transferred from one computer environment to another computer environment.

At this point it is appropriate to discuss two possible disadvantages of the PC environment related to software support and test security. In the software support area, what happens if, for instance, Ashton-Tate no longer supports dBaseIII Plus? What can be done? Usually this problem can be avoided in two ways: The files can be written into other systems via an unformatted ASCII file or the upgrade to dBaseIII Plus (dBaseIV) could be implemented using the utility translator. The security issue can be easily resolved by keeping a particular room 'secured' if the PC has an internal hard disk, or by locking up an external hard disk in a fireproof secure cabinet. In addition, hardware can be used to bolt microcomputers in place.

In summary, the optimal situation would be the development of a relational item bank which links item characteristics to item statistics using, for example, dBaseIII Plus. This item bank (Phase I) would be created using microcomputer software that is portable and could be transferred to a mainframe if desired. With Phase I completed, the linear program packages (possibly the zero-one program) would need to be incorporated into the system. Again, the specific mathematical models to be implemented for Phase II are contingent on the outcomes of future feasibility studies and the decision as to whether or not to have the test developer interact with the system. Therefore, it is recommended that for the present, the item selection and test development

process be accomplished using an automated item banking system. When interactively selecting an item for inclusion on a test form, the test developer would be able to view the Item Information Curve as well as the Test Information Curve with the selected item and the Test Information Curve without the selected item. The automated system allows the test constructor to view the effect an item will have on the characteristics of the total test.

Next, a document publishing system is necessary for the completion of Phase III. In this environment, test items would need to be viewed intact. That is, mechanical rub-ons, illustrations, and item numbers would be stored with the corresponding items. From this system, test booklets could be published with the desired spacing, format, and pitch. If changes are warranted, the corrections could be easily made with the help of a stylesheet component, without violation of quality control standards. Such a system, therefore, would address many of the quality control problems that have plagued test publishing and allow for the successful completion of Phase III.

## REFERENCES


Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Company.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores (pp. 397-479). Reading, MA: Addison-Wesley Publising Company, Inc.

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

Hambleton, R. K., & Swaminathan, A. (1985). Item response theory: Principles and applications (pp. 255-279). Boston, MA: Kluwer Nijhoff Publishing.

Lee, W. M., & Fairbank, B. A. (1983, April). Item bank enhancement. (Available from Operational Technologies Corporation, 5825 Callaghan Road, Suite 225, San Antonio, TX 78228).

Linn, R. L., Hastings, C. N., Hu, P. G., & Ryan, K. E. (1987, April). Armed Services Vocational Aptitude Battery: Differential item functioning on the high school form (AFHRL-TR-87-45, AD-A193 693). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Lord, F. M. (1975). The 'ability' scale in item characteristic curve theory. Psychometrika, 44, 205-217.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

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

Rasch, G. (1966). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Reckase, M. D., & McKinley, R. L. (1983). Some latent trait theory in a multidimensional latent space. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology.

Ree, M. J. (1978). Automated test item banking (AFHRL-TR-78-13, AD-A054 626). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 39, 111-121.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology.

Thorndike, R. L. (1971). Educational measurement. Washington, DC: American Council on Education.

Van der Linden, W. J. (Ed.). (1987). IRT-based test construction (RR 872). Enschede, The Netherlands: Twente University of Technology.

Warm, T. A. (1978). A primer of item response theory (CG941278). Oklahoma City, OK: U. S. Coast Guard Institute.

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