ED 204 349 TM 810 088

AUTHOR Kreitzberg, Charles B.: Jones, Douglas H.
TITLE An Empirical Study of the Broad Range Tailored Test of Verbal Ability.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-80-5
PUB DATE May 80
NOTE 250p.

EDRS PRICE MF01/PC10 Plus Postage.
DESCRIPTORS *Computer Assisted Testing: Higher Education: High School Students: Intermediate Grades: Item Banks: Latent Trait Theory: *Scoring Formulas: Secondary Education: *Student Attitudes: *Test Reliability: *Verbal Ability
IDENTIFIERS *Adaptive Testing: *Broad Range Tailored Test of Verbal Ability: Preliminary Scholastic Aptitude Test

ABSTRACT

The Broad-Range Tailored Test (BRTT) is a computerized adaptive test. Each testee responds to 25 items: at the conclusion of the test the computer calculates a verbal ability score for the individual. The test was designed to yield a verbal ability score from the fifth grade level to the graduate school level. Two forms of the BRTT were administered to 146 high school students. Analyses revealed that the BRTT was more reliable than the Preliminary Scholastic Aptitude Test (PSAT) for a number of scores derived from the data, confirming theoretical expectations regarding the increased efficiency of adaptive, as compared to conventional tests. Of nine observed scores and score transformations, the most useful score was the expected proportion correct over the entire item pool. The accuracy of the BRTT was in accord with theoretical expectation. Student response to the computerized testing procedure was generally quite favorable. Students found the human-computer interface easy to use and less fatiguing than a long pencil-and-paper test. These results suggest that computerized adaptive testing is ready to take the first steps out of the laboratory environment and find its place in the educational community. (Author/BW)

AN EMPIRICAL STUDY OF THE BROAD RANGE TAILORED TEST OF VERBAL ABILITY

BY

Charles B. Kreitzberg and Douglas H. Jones

May, 1980

-FINAL REPORT-

Educational Testing Service
Princeton, New Jersey

ABSTRACT

This document is the final report of a study designed to investigate the performance of the Broad-Range Tailored Test of Verbal Ability. The Broad-Range Tailored Test (BRTT) is a computerized adaptive test developed by Frederic M. Lord. It employs a maximum likelihood selection strategy to choose items from an item pool stored on magnetic disk. The items selected are tailored to the individual testee and are presented on a computer terminal. Each testee responds to 25 items; at the conclusion of the test the computer calculates a verbal ability score for the individual. The test was designed to yield a verbal ability score from the fifth grade level to the graduate school level.

Performance of the BRTT had been investigated by means of simulation studies. The current study is the first empirical test of its performance. Two forms of the BRTT were administered to 146 high school students. The students also answered a posttest questionnaire in which they indicated their reactions to this form of testing.

Analyses revealed that the BRTT was more reliable than the PSAT for a number of scores derived from the data. The test-retest reliability of the BRTT was .8719 at the 25th item; reliability of the PSAT verbal score (scaled down to 25 items) was .65. Analyses of the reliability of the BRTT vs. the PSAT revealed that the tailored test was also more reliable than the conventional test at shorter lengths. Correlations between scores on the BRTT and PSAT were reasonably high--typically about .86. This finding confirms theoretical expectation regarding the increased efficiency of adaptive. as compared to conventional tests.

The study investigated nine of observed scores and score transformations. The most useful of these was found to be the expected proportion correct over the entire item pool. This score was highly reliable and was found to be parallel with respect to the mean values across forms A and B. θ, the commonly-employed latent-trait parameter and [illegible], a monotone transformation of θ, did not exhibit this characteristic.

The information functions of the BRTT were calculated and compared favorably with simulation results previously reported by Lord. Thus the accuracy of the BRTT was in accord with theoretical expectation.

Student response to the computerized testing procedure was generally quite favorable. Students found the human-computer interface easy to use and less fatiguing than a long pencil-and-paper test.

Operationally, the system performed well. Detailed analysis of 11 anomalous cases, suggested refinements to the system. Response time was adequate and consistent. Reliability of the hardware and software was excellent. These results suggest that computerized adaptive testing is

ready to take the first steps out of the laboratory environment and find its place in the educational community.

The recommendations emerging from this study are: (1) that the organization collaborate with an interested client to develop an adaptive test for use in an educational setting; (2) that the potential for microprocessor-based systems for the delivery of adaptive testing be evaluated; (3) that extensions to item response theory and the development of alternative models for the provision of adaptive testing be explored; and (4) that high priority be accorded the development of innovative assessment strategies for computer presentation. Such items might involve simulation and gaming, constructed responses, graphics, motion, sound, and time-dependent responses.

Acknowledgments

This document is the final report of a project designed to evaluate the performance of the Broad-Range Tailored Test of Verbal Ability. The theoretical foundations for this study were erected by Frederic Lord, who designed the tests, investigated their performance in simulation studies, and served as consultant throughout the project.

The implementation of the PDP-11 computer system, whose reliable performance was a source of constant reassurance during the data collection phases of the project, was largely the work of Martha Stocking and Len Swanson. Len served as co-project director during the first phase of the project, and Martha has continued to actively support the project despite a change in organizational locus. Martha also reviewed drafts of this report, re-analyzed anomalous case data, and helped untangle some particularly knotty problems. Len and Martha's contributions will be found throughout this report.

The field trials were organized, conducted, and monitored by Lisa A. Morgan, field researcher for the study. Lisa administered the test instruments as well as collected and assembled the data. The field trials could not have been conducted without the support of Susan Wood. Susan provided technical support for both hardware and software; her more spectacular interventions reduced a number of potential crises to minor problems.

Two local high schools demonstrated their concern with and interest in educational research by permitting us to invade their space, disrupt their schedules, and recruit their students. Without them, this study could not have taken place. Special thanks are due to Mr. George Petrillo, then principal of Princeton High School; Ms Florence Burke, assistant

principal; and Mr. Robert Gorman, who was then Director of Guidance at Princeton High School and as deeply supportive as anyone could be.

Mr. Richard Kaye, principal, and Dr. Ruth Samsel, Director of Guidance, welcomed us to South Brunswick High School. Mr. Joseph Valenti, mathematics teacher at South Brunswick High School, involved himself in every aspect of the data collection and created a well-organized administrative system to support our efforts.

Carolyn Roth excellently programmed the statistical analyses. Bill Van Hassel performed some heroic manipulations to produce a preliminary analysis of the data for presentation at the 1979 meeting of the International Association for Educational Assessment.

Thanks are certainly due to Audrey Brown, Jill Callahan, Kathy Fairall, Barbara Goldberg, Elaine Guennel, Charity Hicks, Peggy Jackson, and Christine Sansone, whose collaborative efforts saw this manuscript through numerous drafts. Elsa Rosenthal smoothed out our uneven and inconsistent verbiage. Meg Monroe performed a literature search which was used in Chapter IV.

Winton H. Manning headed-up the ETS management team that supported the project. Ernie Anastasio, Al Beaton, Sam Messick, Steve Stupak, and Bruce Taylor provided administrative support and review. Special thanks is due to Bruce Taylor whose sympathetic ear and counsel often calmed a sea of confusion.

Finally, thanks must go to the 146 high school students who were willing to devote their free periods to our study even though we asked them to take a test.

Charles B. Kreitzberg

Douglas H. Jones

TABLE OF CONTENTS[1]

Appendices

[1] The authors would like to acknowledge that portions of the content of Chapters I and II are drawn from work previously published collaboratively with Martha Stocking and Len Swanson and referenced therein. Martha and Len were also major architects of the computer system used in this study.

LIST OF FIGURES

LIST OF FIGURES (CONT'D)

LIST OF FIGURES (CONT'D)

LIST OF FIGURES (CONT'D)

LIST OF TABLES

LIST OF TABLES (CONT'D)

Chapter I

Background of the Study

1.1 Purpose of the Study

As a major testing organization, Educational Testing Service has a long-standing interest in and commitment to improving the testing process. Although the organization uses computer technology to support the administrative aspects of its testing programs (such as candidate registration, item analysis, scoring, and reporting), there has been little use of the computer as a vehicle for presenting test items.

Although the prospect of employing computers as testing devices has intrigued psychometricians for over a decade (Weiss, 1973), two considerations have militated against computerized testing: the first obstacle was the high cost associated with this technology; the second was the lack of an adequate psychometric theory to support adaptive testing. In recent years, both obstacles have become less problematic.

The development of microelectronic technology has radically reduced the cost of computer hardware, and forecasts predict this trend to continue for a number of years (Noyce, 1977). It seems likely that computers will soon be as readily accessible as telephones.

The development of item response theory (Lord and Novick, 1968) provided a psychometric foundation on which adaptive tests could be erected. A number of investigators have developed adaptive testing models and explored their performance in both simulation and empirical studies (McBride, 1976). The convergence of psychometric and technical developments suggests the feasibility of practical computerized adaptive testing.

One of the most promising of the adaptive testing models recently ceveloped is Lord's (1977a) Broad-Range Tailored Test of Verbal Ability (BRTT).

The Broad-Range Tailored Test of Verbal Ability employs an item pool stratified into difficulty levels and arranged by item type within difficulty. The BRTT yields an ability score appropriate to students from the fifth grade level to graduate school. McBride has characterized the BRTT as "the most ambitious adaptive testing proposal to appear in the literature, by virtue of the range of ability over which Lord intended it to be used (McBride, 1976, p. 54)".

In designing the BRTT, Lord investigated about thirty designs for a broad-range tailored test administering each to approximately 1000 simulated examinees. The final design is described in detail in section 2.1.

Lord (1977a) suggested that the appropriate next step in the evaluation of the BRTT would be an empirical study of its performance when administered to actual (rather than simulated) examinees. The present study was designed to explore the empirical performance of the BRTT.

The present study involved two phases. The first was to design and implement a computer system capable of administering the BRTT. The second phase involved administering the two forms of the BRTT to 146 high school students. The students' responses were analyzed to determine their relationship to theoretical expectations.

1.2 Historical Antecedents of Computerized Adaptive Testing

Weiss (1976) has traced the history of adaptive testing to Binet's procedures for the assessment of intellectual functioning. Binet's procedures were characterized by three aspects that are typical of contemporary adaptive tests:

1. Prior information is used to select a starting point for the assessment procedure. The tester determines the initial item based on his or her judgment of the testee's ability.
2. The items presented depend, in part, upon the testee's responses to previous items. Basal and ceiling levels are used to ensure that most items are within the appropriate range of ''fficulty and are neither too difficult nor too easy for the individual.
3. A stopping rule is employed to determine the point at which the administration of items ceases. Thus, individual testees may receive tests of different lengths.

Although individually administered tests may be subject to some distortion owing to testee-examiner interaction effects (Rosenthal, 1966; Wickes, 1956), this disadvantage is balanced by the examiner's ability to maintain rapport, clarify ambiguities in items or responses, record response latencies, probe responses of interest, and, generally, manipulate the conditions of test administration to obtain information yield. Thus, the individually administered test has the potential to elicit considerable information about the examinee. Because of their high yield, these tests are often employed as clinical instruments (Harrison, 1965). Unfortunately, individualized administration is too costly and time-consuming a process to be employed in many assessment situations.

The inefficiency of individually administered testing has created a need to adopt less time-consuming and less expensive methods when large numbers of persons are to be tested. The group-administered standardized paper-and-pencil test has become the accepted compromise between the desirability of individual assessment and the need for efficient testing of large numbers of people (McBride, 1976).

The need for standardized group-administered tests was recognized prior to World War I. By 1918, the need for rapid classification of recruits had spurred the development of the Army Alpha and Army Beta tests, initiating a period of tremendous growth in group testing (Weiss and Betz, 1973).

Group tests have a number of important advantages over individual tests. Among these are the following:

1. Lower cost to administer and score since a number of individuals can be tested by a single administrator and machine scoring of answer sheets is possible.
2. Reduction in examiner-effect variables due to reduction in relationship factors; less reliance on examiner's judgment in scoring.
3. Comparisons among testees are facilitated because each individual receives the same set of items under uniform conditions.

Despite its economic and procedural advantages, the group-administered objective multiple-choice test is far from an ideal testing instrument, and psychologists have been intrigued by the potential utility of assessment procedures that adapt to the individual.

Hutt (1947) compared "consecutive" with "adaptive" administration of the Stanford-Binet. In the adaptive technique, he administered an easier item following an incorrect response and a more difficult item following a correct response. Hutt found that students who had poor school adjustment obtained reliably higher IQ scores in the adaptive modality.

Hick (1951) employeo Shannon's model of information to develop an "up-down" technique in which testees would receive items for which they had a 50% probability of choosing the correct answer. This procedure was intended to obtain maximum information from the item responses. The notion of information yield is a central one in current adaptive testing strategies.

One form of adaptive test which can be administered with paper and pencil is that in which several "peaked" tests are created, each with overlapping ranges. An examinee who scores in the extreme range of a test is retested with a more appropriate instrument.

One strategy, called the two-stage adaptive test, involves administering a short "routing test" to all examinees, who are then directed to an appropriate second-stage test with items of relatively homogeneous difficulty (McBride, 1976).

A two-stage adaptive strategy was tried experimentally by Angoff and Huddleston (1958), who concluded that although the use of two narrow-range (peaked) tests was slightly more reliable than a single broad-range test (.89 vs .85), the increase in validity coefficient for the two-stage procedure would not exceed .02 on the average. One problem with a two-stage adaptive strategy is that errors of measurement would cause some testees to be misclassified by the routing test and to be routed to an inappropriate second-stage test. Angoff and Huddleston felt that the numbers of such misclassified students, although small, would be sufficient to cause serious administrative problems and that the advantage of heightened reliability "would not be great enough to warrant changing to the administratively more complex two-level test system" (Angoff and Huddleston, 1958, p. 4).

This strategy has also been investigated by Betz and Weiss (1974) and Vale (1975).

Lord (1971) proposed a paper-and-pencil branching test which he termed a flexilevel test. A flexilevel test of length k contains 2k - 1 items. Testees begin at the item of middle difficulty and are branched to an item of greater difficulty following a correct response or to an easier item following an incorrect response.

There is some evidence that paper-and-pencil adaptive tests are superior to conventional tests of the same length (McBride, 1976; Vale, 1975). However, most research on adaptive testing has focused on branching strategies which are sufficiently complex to require computer administration.

A more sophisticated alternative to paper-and-pencil adaptive testing is to employ a computer to select and administer individual items. Testing in which the computer is used to individually select items has been variously referred to as adaptive testing (Weiss and Betz, 1973), programmed testing (Cleary, Linn, and Rock, 1968a), branching tests (Bayroff and Seeley, 1967), response-contingent testing (Wood, 1973), tailored testing (Lord, 1970), and computerized adaptive testing (Kreitzberg, Stocking, and Swanson, 1978).

Computerized adaptive testing (CAT) has recently been a subject of considerable research (Weiss and Betz, 1973; McBride, 1976). Conferences on adaptive testing were held in Washington, D.C. in 1975, and at the University of Minnesota in 1977, and 1979. In addition to the active program being conducted at the University of Minnesota (Weiss, 1975), the U. S. Civil Service Commission has been conducting research with a view toward automation of Civil Service testing (Urry, 1976). Researchers at

Minnesota and the U. S. Civil Service Commission have developed computer systems capable of administering adaptive tests.

The variables that have been studied by researchers include: reliability and validity (Weiss, 1973; Waters, 1974; Urry, 1976), accuracy at extremes (Lord, 1970), ability to reproduce conventional test scores (Linn, Rock and Cleary, 1972), information yield (Lord, 1970), effects of varying step sizes (cf. Wood, 1973; McBride, 1976), measurement error (Wood, 1973), and number of items presented (Wood, 1973; McBride, 1976). Empirical research concerned with comparisons to conventional tests has focused on: external validity (Olivier, 1974; Waters, 1974), internal consistency reliability (Vale and Weiss, 1975), test-retest temporal stability (Betz and Weiss, 1974; Larkin and Weiss, 1974), and characteristics of score distributions (Betz and Weiss, 1973).

In section 1.5, major strategies for computer-administered adaptive testing will be considered. Prior to discussing computerized adaptive testing, it is appropriate to review the psychometric foundations on which computerized adaptive tests are built.

1.3 Psychometric Foundations

In classical test theory (Gulliksen, 1950), item parameters are defined in terms of group data. For example, the difficulty of an item is defined as the proportion of individuals who get the item correct. This proportion is not an inherent property of the item, but will vary with the group; a given item may be difficult for a group of low ability and easy for a group of high ability.

Although it is possible to develop adaptive tests based on classical test theory (e.g. Angoff and Huddleston, 1958) there are three issues which are not easily resolved within this theory (Kreitzberg, Stocking, and Swanson, 1978):

1. Scoring. Since different examinees receive different items, the traditional number-right score used by classical test theory is inappropriate. This raises questions regarding the method of scoring the test and the comparison of scores received by different individuals.
2. Item parameters. Since appropriate items are selected individually for each examinee, item characteristics must be population-invariant. However, as previously noted, classical test theory employs group-dependent item parameters.
3. Comparing strategies. There are many possible strategies for selecting items and scoring responses. Conventional tests are usually evaluated by such measures as reliability and validity. As these correlational indices are group-dependent, they may not be appropriate for adaptive tests, since adaptive testing requires that methods of comparing different strategies and scoring procedures be independent of the group taking the test.

Unlike classical test theory, item response theory (Lord and Novick, 1968) allows the test scores of all examinees to be expressed within a common metric, regardless of the fact that each examinee may have answered different, and even different numbers of, items. This metric allows ordering of examinees with respect to the trait to be measured and quantification of the magnitude of the differences among examinees. The item parameters employed by item response theory are independent of the group to which the item is administered. In addition, techniques for comparing item selection strategies and scoring procedures have been developed which involve a consideration of the amount of "information" obtained from a test at various levels of the trait being measured (Lord and Novick, 1968), and which are also independent of the group to which the test is administered.

In item response theory, it is assumed that some underlying trait is to be measured. As this trait is unobservable, it is often called a latent trait. It is assumed that the latent trait is unidimensional. Generally, ability traits are considered although achievement and personality traits may also fit the model.

An individual is considered to possess a score θ which indicates the level of trait he or she possesses. The true score of classical test theory is an isomorphic transform of θ (Lord, 1980).

In item response theory, the probability of a correct response to an item is assumed to be a function of the individual's ability level, θ, and the psychometric properties of the item. For every possible level of θ, there exists a probability of a correct response. The graph of this function is typically shaped like a normal ogive; the exact shape of the curve depends upon the psychometric properties of the specific item and on the model chosen. Figure 1.3a illustrates some typical item characteristic curves.

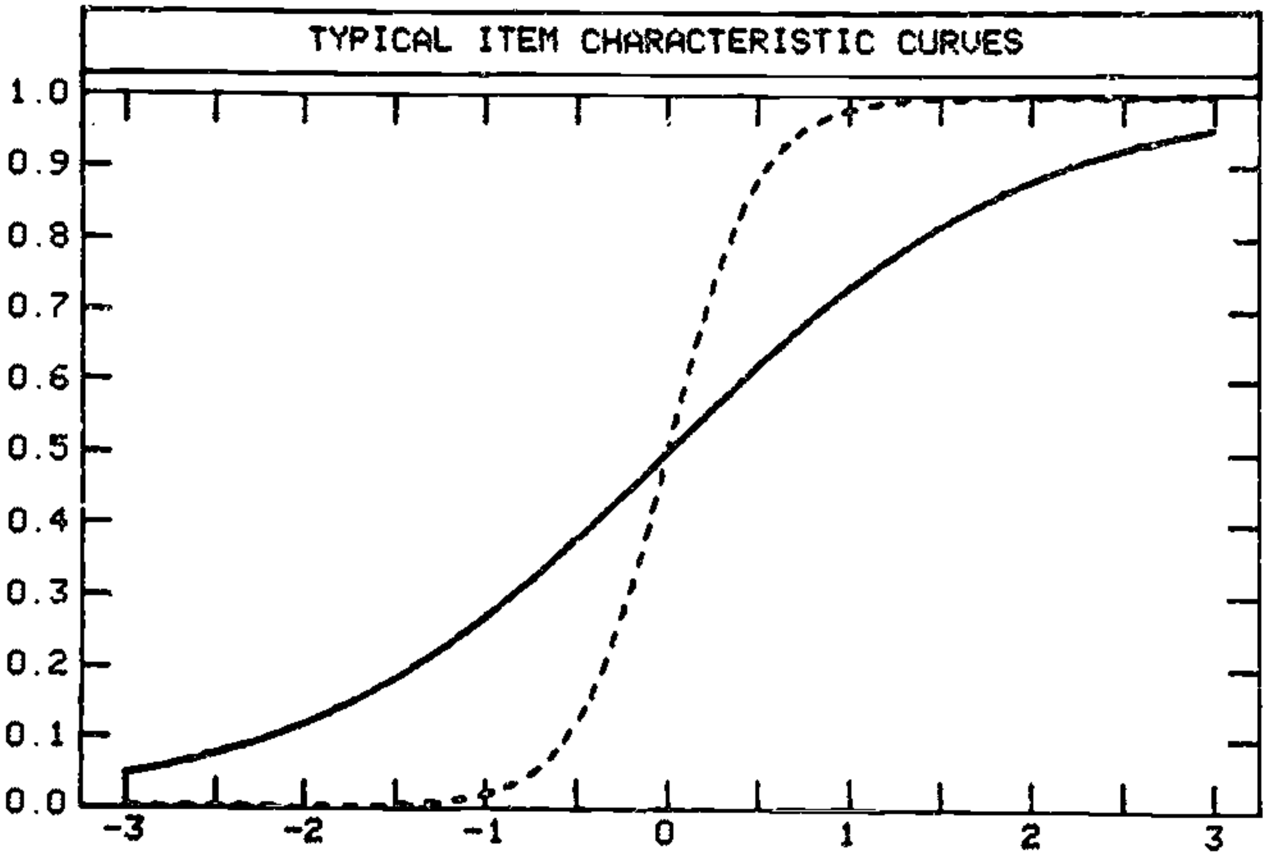


Figure 1.3a Typical Item Characteristic Curves

Because the normal ogive is mathematically intractable, an alternative model known as the logistic model is generally employed. With the logistic model, the probability of a correct response, given θ, is:

$$P(\theta) = \left[1 \quad e^{1.7a(\theta - b)}\right]^{-1}$$

(Lord and Novick, 1968, p. 400).

The logistic model is characterized by two parameters _a_ and _b_. Individuals with very low values of θ may sometimes get an item correct by chance. To account for this guessing factor, a third parameter _c_ is added to the model (see section 2.3). The three-parameter logistic model is the theoretical parent of the Broad-Range Tailored Test investigated in the current study.

If θ is the trait parameter and x is a general test score function of the response vector, then the information function $I_x(\theta)$ is defined as:

$$I_x(\theta) = \frac{\frac{\partial}{\partial\theta} E(x|\theta)}{\sigma^2_{x|\theta}}$$

The information function is a useful tool in evaluating the performance of a test.

Since the classical notions of reliability are insufficient for latent trait theory, item-selection strategies and ability-estimation procedures are often compared through the use of information functions

(Lord and Novick, 1968). While the concept of an information functicn is mathematically precise, its properties have great intuitive appeal.

In particular, for appropriate models of $P(\theta)$, the maximum likelihood estimate θ has an information function inversely proportional to the length of the confidence interval for estimating the ability parameter θ. The higher the information function, the more precise the estimate of ability. By comparing information functions of tests, it is possible to determine which test yields the greatest precision of measurement at different levels of ability.

The information function for a conventional test using the maximum likelihood estimate of θ is proportional to the number of items in the test (Lord and Novick, 1968). This allows any comparison between information functions for a conventional test and a tailored test to be discussed in terms of the number of items that must be added to or deleted from the conventional test to obtain the same amount of information available from a tailored test, at various ability levels.

An excellent review of the material in this section will be found in Sympson (1977).

1.4 Rationale for Adaptive Testing

Paper-and-pencil tests are generally designed to measure a reasonably wide range of abilities. Consequently, such tests include a range of item difficulties to permit them to discriminate among a diverse population of testees. Unfortunately, the need to restrict the test to a reasonable length results in fewer items at the extremes than in the middle of the range. This restriction means that conventional tests are most precise at the center of their range of measurement, and precision of

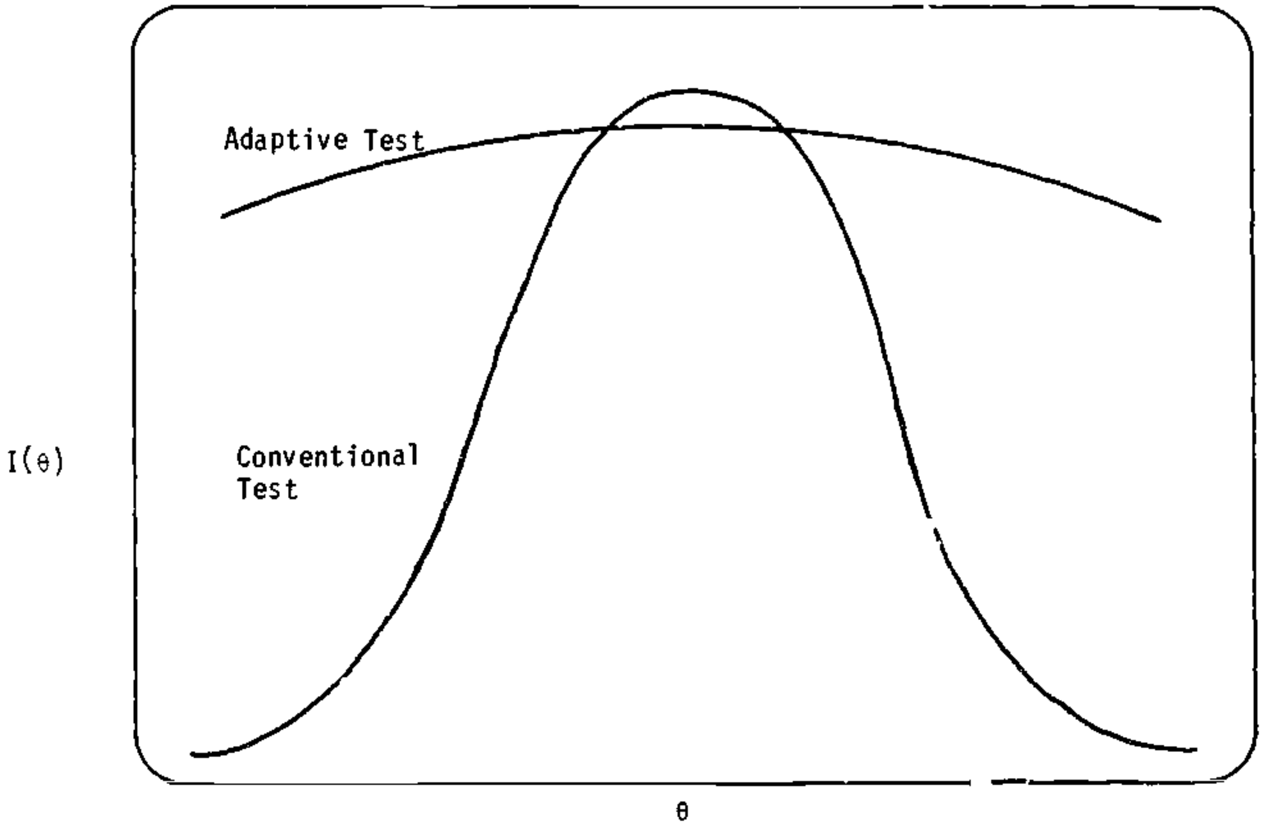


Figure 1.4a Hypothetical Information Curves from a Conventional and Adaptive Test

measurement declines toward the extremes.

Ideally, a test would be "tailored" to an individual and would comprise test items that are clustered around the individual's ability level. The more closely a test approximates this ideal, the greater will be its precision of measurement. Computerized adaptive testing employs iterative techniques to select items which cluster around the individual testee's ability level. Although various adaptive stracegies exist (see section 1.5), they generally follow a similar pattern:

1) An initial estimate of the testee's ability level is made in some convenient way (e.g., grade level, age).
2) The ability estimate is used to select an appropriate item from the item pool.
3) The item is scored correct or incorrect, and an estimate of the testee's ability level is calculated.
4) If the estimate is sufficiently precise, the procedure is terminated. Otherwise the estimate is further refined by returning to step 2.

Kreitzberg, Stocking, and Swanson (1978) have recently reviewed the status of computerized adaptive testing and have enumerated its potential advantages in terms of its properties.

Perhaps the major advantage of adaptive testing is that, in general, fewer items are required to achieve a specified level of measurement accuracy than are required in a conventional test. Numerous research studies (cf. Lord, 1970) have confirmed this. The increased efficiency of an adaptive test occurs because the most information is obtained about an examinee if the items administered have a 50% probability of being answered correctly--65% of the time if guessing is taken into consideration (for a five-choice item) (McBride, 1976). Items which are too easy or too hard for a given individual contribute little information about the examinee (Sympson, 1977). Since the purpose of adaptive testing is to choose and administer those items which contribute most to the estimate of an individual's ability level, fewer items are required to achieve the same level of measurement precision. The information function of an adaptive test is higher at any point on the ability scale than that of a conventional unpeaked test, and higher at the extremes than a conventional peaked test. It is also less variable throughout the ability range (Lord, 1977b).

Improvements in measurement precision have been established theoretically and verified in simulation studies. The amount of improvement to be expected with a given test depends on the size and characteristics of the item pool. As an example, Urry (1976) suggests a roughly five-to-one (80%) reduction in the number of items required to achieve reliabilities comparable to conventional test scores.

As a consequence of its higher and less variable information function throughout the ability range, an adaptive test is particularly superior to a conventional test at the extremes of ability. This situation is depicted in Figure 1.4a. The wider the range of ability being measured, the greater this discrepancy. Since underlying ability is not usually directly measureable, this result cannot be verified empirically; however, it has been demonstrated theoretically (Lord, 1977a). It is a particularly important advantage with respect to testing lower-ability students, since in a conventional test the accuracy of such measurements is virtually swamped by random error introduced by guessing.

Another consequence of the higher and less variable information function of an adaptive test is that scores better reflect the true distribution of ability in a population. Weiss (1975) has demonstrated this in simulation experiments. This is important when group, as well as individual, characteristics are of interest.

The latent trait theory underpinning adaptive testing contributes another important advantage: scores based on latent trait theory are on an interval scale. McBride (1976) notes that scores based on classical test theory are on an ordinal scale. Thus the magnitude of differences between scores has a natural meaning in latent trait theory, but not in classical test theory.

There is some evidence to suggest that scores on adaptive tests have greater temporal stability (test-retest reliability). Weiss (1973) cites results of live testing experiments that indicate this, and claims that simulation studies show that it holds over the entire ability range.

Finally, computerized adaptive testing may reduce some of the random error in conventional tests owed to confounding of power conditions. It has often been noted that, because of administrative requirements, some element of speededness is frequently introduced into group-administered power tests. Weiss (1973) cites data showing that speededness differentially affects individuals, thus confounding the predictive qualities of the test. This problem can be virtually eliminated with computerized adaptive testing, since administration of the test is individualized, and time limits can be controlled by the examiner.

Administrative Effects

A great deal of attention has been given in both group and individually administered testing to standardization of the testing environment, control of administrator effects, objectivity of scoring, and security of materials. Computerized adaptive testing offers potential advantages over conventional testing in all of these areas.

Urry (1975) points out that computerized testing is more standardized than conventional testing because the administrative procedures are programmed and, therefore, more uniform and controlled. This reduces differential effects of the testing environment. In individually administered tests, studies have shown that administrator effects and clerical errors in scoring may seriously compromise test objectivity (Weiss, 1973). For example, factors such as expectancy, knowledge of the testee, degree of rapport, and race have all been shown to influence individual scores. Even in group administrations, the examiner may induce different levels of stress in different individuals (Weiss, 1973). Since computerized testing eliminates the human examiner and precisely controls administration and scoring, these effects should be better controlled, if not eliminated.

Characteristics of the answer sheet and item arrangement have also been shown to affect group scores, as well as differentially affect individual scores (Weiss, 1973). This compromises the psychometric qualities of the test. In computerized testing these effects are eliminated. They are, however, replaced by a new set of factors relating to the interface between the individual and the testing device. Because of the relative newness of the field, these factors have been largely unexplored. However, computerized testing provides a degree of control impossible in a conventional testing environment. This should make it practical to easily modify the testing environment as new evidence regarding the effects of that environment is uncovered.

Computerized testing should make it easier to safeguard the security of test materials. It has been argued (Wood, 1973) that, since test booklets are no longer needed, and since different individuals receive different items, the security problem will be diminished. This assumes that adequate procedures to safeguard the integrity and accessibility of the computer have been developed. As computer security systems continue to improve, this should be the case.

Computerized testing provides significant advantages in the scheduling of test administrations. Tests can be administered at times and locations convenient to the student. For example, walk-in test centers become possible. Even if test security requires that all administrations be simultaneous, it may be possible to locate terminals at the convenience of the student and virtually eliminate the administrative procedures involved in registration and arrangements for test centers.

Finally, many of the other administrative procedures required for conventional group-administered testing can be reduced or eliminated with computerized testing. The list includes: test booklet and answer sheet printing, storage and distribution; accounting for and return of materials; answer sheet processing; certain aspects of score reporting; and test center management. These administrative procedures are, of course, replaced by others required for computer administration of tests. However, the latter should, once established, be simpler and less costly to carry out on a routine basis.

Affective Factors

Little research has been conducted to date on the affective implications of computerized adaptive testing. Some researchers have hypothesized several advantages of computerized testing in terms of its effects on the testee. Chief among them is that it may increase the student's interest in and motivation for taking the test. Johnson and Mihal (1973) report results that showed that blacks perform better on a computerized test, and suggest that motivational and examiner effects might have been responsible. Weiss (1975) found similar effects when feedback on the correctness of a response is provided. These results suggest that the computerized testing environment may in some cases be more motivating or less anxiety-producing than the conventional testing environment.

It has also been hypothesized that, because items better match the ability level of the testee, adaptive testing may have a positive effect on the attitudes of high- and low-ability students. Weiss (1975) suggests that the high-ability student may be bored by a conventional test, with a resulting deterioration in performance. The low-ability student may be similarly affected by the frustration and anxiety resulting from attempting items that are overly difficult. In addition, there is evidence that low-ability students guess more frequently, thus introducing greater error into the score (Weiss, 1973). Computerized adaptive testing should tend to reduce these negative factors.

Limitations of the Group-Administered, Multiple-Choice Mode of Testing

Computerized adaptive testing potentially offers several other advantages over conventional testing methods. These advantages result from the power and flexibility inherent in computer administration of a test. One advantage is the ability to gather and report additional information about the testing process that cannot readily be gathered in conventional paper-and-pencil testing. For example, Weiss (1975) developed a measure he refers to as a "consistency" index on each student. This measure is, roughly, the number of strata or difficulty levels administered to the student. Weiss showed that this measure is generally correlated to the test-retest reliability of the student's score. If this is so, then reporting this measure may provide additional information helpful in evaluating the student's score.

Another example of additional information that can be gathered is response latency--the time it takes a student to answer an item. Green (1970) suggested that response latency may be related to guessing. Additional research will be needed to determine the value, if any, of latency information.

Computerized adaptive testing provides an opportunity for greater flexibility in the testing process itself. For example, students can be given feedback on the correctness of their responses. Weiss (1975) showed that black students tend to score better and omit fewer items when feedback is provided. Students can also be permitted to re-try an item after an incorrect response. This may be useful in analyzing or weighing wrong answers. Finally, information about the testing session, including the student's score(s), can be provided immediately. This may be beneficial to the users of test scores, as well as to the students themselves.

Computerized testing more readily permits alternatives to the multiple-choice item type than does conventional testing. These alternatives might include free or constructed response items, and probabilistic response items (ones in which the student assigns weights or priorities to the choices). While such alternatives can be done with paper and pencil, they are difficult to administer and score. Computerized testing may, therefore, open up new approaches to item design.

1.5 Strategies of Adaptive Testing

This section provides a brief summary of some of the major strategies which have been employed for selecting items in an adaptive test. Weiss (1974) and McBride (1976) have published extensive reviews of adaptive-testing strategies.

Two-stage

One of the simplest adaptive-testing strategies, the two-stage strategy, has been previousl' discussed. Typically, the two-stage strategy employs a routing test which provides an initial estimate of the individual's ability. Based on the score attained on the routing test, the individual is then branched to a measurement test appropriate to his or her ability level. A major advantage of the two-stage strategy is that it can be used with paper-and-pencil testing provided that it is administratively feasible to score the routing test before providing the examinee with the second-stage test. The major disadvantage of this strategy is that misrouting due to error has serious measurement consequences.

Pyramidal

Pyramidal models employ items which are structured in a tree as illustrated in Figure 1.5a. The testee begins at the top of the pyramid and is administered the initial item which is then scored correct or incorrect. Following an incorrect response, the testee is branched to an item of lower difficulty; following a correct response, the testee is branched to an item of higher difficulty. Lord's (1971) flexilevel item strategy, discussed previously, is a variant of a pyramidal branching model. Many other variations are possible and have been studied. Pyramidal models are somewhat sensitive to errors of measurement, as a correct guess or an incorrect response due to a confounding variable (for example, failing to understand a key word in a nonverbal item) may affect the reliability of the final score. Weiss (1974) has pointed out that the pyramidal model does not guarantee that items will cluster at the 50 percent probability of a correct response, as desired for maximum information yield.

Stradaptive

The stradaptive strategy developed by Weiss (1973) employs an item pool which is divided into strata based on item difficulty. The initial item is selected on the basis of a prior estimate of the individual's ability, and branching occurs between strata. The stradaptive strategy differs from the pyramidal strategy in that branching is from stratum

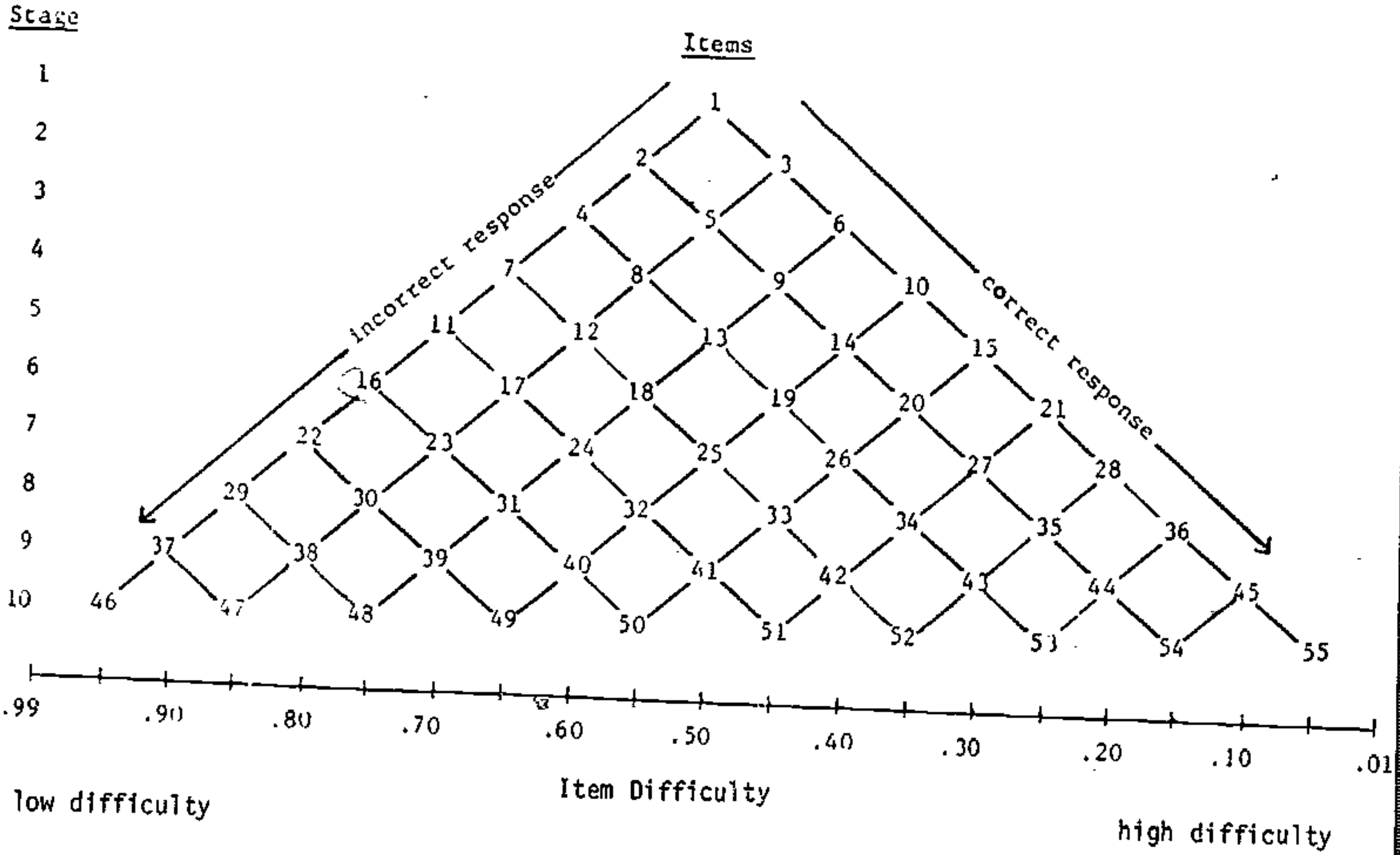


Figure 1.5a A Pyramidal Item Structure (Figure adapted from Weiss, 1974)

Reproduced by permission from David J. Weiss, Strategies of adaptive measurement, copyright © 1974. Research report 74-5. Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota.

to stratum rather than from item to item, although the same types of branching rules may be employed in both.

Bayesian

A number of investigators have explored Bayesian models for item selection. The general procedure used in the Bayesian model is to obtain a prior estimate of the testee's ability along with an estimate of the uncertainty. The item in the pool selected is the one which will most reduce the uncertainty of the ability estimate. Following administration of the item, the prior ability and uncertainty estimates are revised to yield a posterior ability estimate. This posterior estimate replaces the prior estimate for the next iteration. The procedure may be continued until a posterior ability estimate has been obtained which has the desired degree of accuracy. Novick (1969) and Owen (1975) have proposed Bayesian procedures for adaptive testing.

Maximum Likelihood

The final class of adaptive testing strategies to be considered are the maximum-likelihood strategies. The maximum-likelihood procedure is similar to the Bayesian procedure, although the mathematics underlying it is quite different. The maximum-likelihood procedure requires that the testee answer at least one item correctly and one incorrectly. When this has been accomplished, the maximum-likelihood equations can be solved to yield an ability estimate and standard error. This estimate is refined iteratively as in the Bayesian procedure. The Broad-Range Tailored Test of Verbal Ability, which was the subject of the current study, employed a maximum-likelihood strategy. The structure of this test is described in detail in Chapter II.

Chapter II

The Broad-Range Tailored Test of Verbal Ability

2.1 Introduction

The Broad-Range Ta'lored Test of Verbal Ability (BRTT) was developed by Lord (1977a). The test employs an item pool stratified into difficulty levels and arranged by item type within difficulty. The computer initially selects specific items by an up-down rule; later items are selected by a maximum-likelihood algorithm. The BRTT yields an ability score appropriate to students from the fifth grade level to graduate school. McBride has characterized the BRTT as "the most ambitious adaptive testing proposal to appear in the literature, by virtue of the range of ability over which Lord intended it to be used (McBride, 1976, p. 54). In designing the BRTT, Lord investigated about thirty designs for a broad-range tailored test, administering each to approximately 1000 simulated examinees. The current study is the first empirical test of its performance.

The BRTT proved quite robust with regard to the selection of the initial item. As shown in Figure 2.1a, Lord (1977a) found little difference in the standard error of measurement at 13 different ability levels related to the difficulty of the initial item.

Lord's simulations included designs whose item matrices contained more or fewer difficulty strata than the 10 columns (shown in Table 2.1b) employed in the present study. He found that a test with the same difficulty range, but employing 363 items stratified into 20 groups, was at least twice as good as the 10-column, 182-item test of Table 2.1a. These results suggest that selection from a 363-item pool would support a much better 25-item test than selecting an equal number of items from

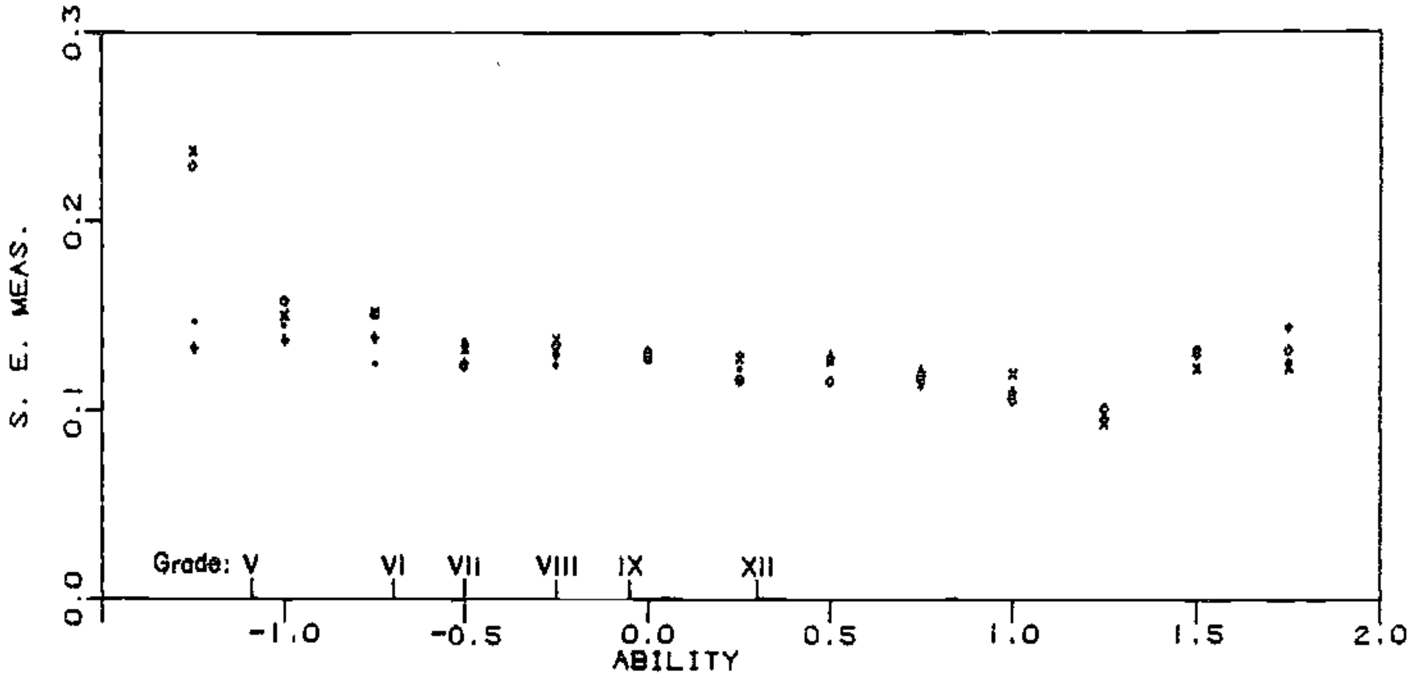


Figure 2.1a The standard error of measurement at 13 different ability levels for four different starting points for the 25-item broad-range tailored test.

Table 2.1b

Broad-Range Verbal Test ltems Arranged by Difficulty Level and Serial Number.

(a,b,c,d,e represent different verbal item types.)

Item Serial No.	(easy) ← Item Difficulty Level → (hard)									
	Grade Level: IV	V	VI	VII VIII	XII					
1			a	a	a	a	a	b		
2			e	e	e	e	c			
3			d	d	d	d	d	d		
4			e	e	e	e	c	c		
5			d	d	d	d	d	d		
6			a	a	a	a	b	b		
7			e	e	e	e	c			
8		d	d	d	d	d	d			
9			e	e	e	c	c	c		
10		d	d	d	d	d	d			
11			a	a	a	a	b	b	b	b
12		e	e	e	c	c	c	c		
13			d	d	d	d	d	d		
14		e	e	e	c	c	c	c	c	
15			d	d	d	d	d	d	d	
16			a	a	a	b	b	b	b	b
17		e	e	c	c	c	c	c	c	
18	d	d	d	d	d	d	d	d		
19		e	e	c	c	c	c	c	c	
20	d	d	d	d	d	d	d	d	d	d
21		a		a	a	b	b	b	b	b
22	e	e	c	c	c	c	c	c	c	c
23		d	d	d	d	d	d	d	d	d
24	e	e	c	c	c	c	c	c	c	c
25		d	d	d	d	d	d	d	d	d

(Table from Lord, 1977)

Reproduced by permission from Applied Psychological Measurement, edited by David J. Weiss, A Broad-Range Tailored Text of Verbal Ability by Lord, copyright © 1977, Volume 1, Number 1, West Publishing Company. All rights reserved.

a smaller, 182-item pool. Still better tests could be produced by using still larger item pools, selecting the best 25 items for each examinee. The item pool size used in the current study was based upon the number of items available from Lord's (1977a) simulations; it was chosen for practical reasons rather than theoretical optimality.

2.2 Comparison of the BRTT to the PSAT

Lord compared the information yield of the BRTT with the Preliminary Scholastic Aptitude Test of the College Entrance Examination Board. Figure 2.2a shows the information function for the verbal score on each of three forms of the PSAT adjusted to a test length of 25 items, compared to the information function for the Verbal score on the Broad-Range Tailored Test. In the tailored test the initial item administered was at a difficulty level appropriate for average college students. The PSAT information functions were computed from estimated item parameters; the tailored test information function was estimated from responses of simulated examinees.

McBride (1976) argued that comparing the BRTT to the PSAT adjusted to a 25-item length may have been unfair since the BRTT selects the 25 "best" items, whereas the PSAT items have divergent discriminating power. He suggested that a preferable comparison of BRTT to PSAT would compare the 25 "best" items of the PSAT, where best is defined as either the 25 items with most discriminating power or the 25 items with the most information at a given ability level. Both McBride and Lord agreed that the results of the simulations were promising and suggested that the procedure be attempted with actual examinees. The current

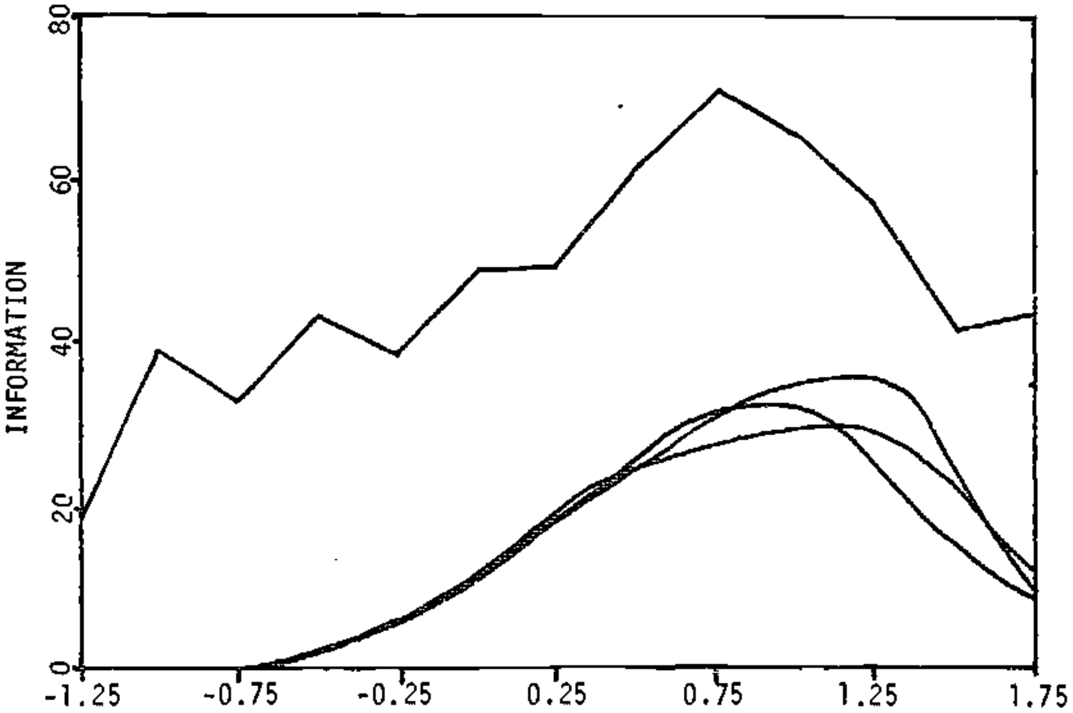


Figure 2.2a Information function for the 25-item tailored test, also for three forms of the Preliminary Scholastic Aptitude Test (lower lines) adjusted to a test length of 25 items.

Reproduced by permission from Applied Psychological Measurement, edit. ' by David J. Weiss, A Broad-Range Tailored Text of Verbal Ability by Lord, copyright © 1977, Volume 1, Number 1, West Publishing Company. All rights reserved.

study solicited PSAT scores from the students participating for purposes of the comparison.

2.3 The Item Pool

The items making up the two forms of the Broad-Range Tailored Test were selected from five ETS-administered testing programs: the Graduate Record Examinations (GRE), the Scholastic Aptitude Test in both standard (SAT) and preliminary (PSAT) forms, the School and College Aptitude Test (SCAT), and the Sequential Tests of Educational Progress (STEP).

An initial item pool consisting of 898 items was obtained by selecting all verbal items that were one of the following item types: (1) synonyms, (2) opposites, (3) incomplete sentences, (4) word relaticns, (5) sentence comprehension. A detailed description of the item pool is provided in Appendix B.

Estimates of the three item parameters were obtained by the LOGIST program. The items were placed on a common scale by obtaining previously computed equatings which related number-right scores among all tests.

The equating of test _s_ to test _p_ was accomplished by employing LOGIST to compute: θ - ability estimates for each person.

a_s, b_s, c_s - item parameters for each item.

Using the 3 paramete. logistic model the probability of a correct answer to item i given θ is expressed by the equation:

$$P_i(\theta_s) = c_i + (1-c_i)/(1 + \exp(-Da_i(\theta_s-b_i))).$$

An estimate of the true number-right score, ξ_s, is given by

$$\hat{\xi}_s = \sum_{i=1}^{n_s} P_i(\theta_s).$$

Similarly, the analogous quantities

$$\theta_p$$

$$a_p, b_p, c_p$$

$$P_i(\theta_p), \text{ and } \hat{\xi}_p = \sum_{i=1}^{n_p} P_i(\theta_p).$$

were computed for test p. The transformation of θ_s to θ_p was then computed using knowledge of the number-right relationship between s and p to place the item parameters on a common scale.

2.4 Structure of the Item Pools

The item pools were stored on disk and were indexed by means of an item-type table. This table was structured as a rectangular array with k rows and m columns. Each column 1,...,m represented a range or stratum of ability (θ). Rows 1,...,k indicated an item sequence; the type of the i^th item administered was specified by row i. Table 2.1b, shown previously, is the item type table form Form A.

The two item-type tables employed in the current study contained 25 rows and 10 columns each and were developed by randomly splitting a 20-column table. The pool for Form A contained 183 items; Form B contained 180 items (Swanson and Stocking, 1977).

The item-type table determined the type of item to be administered at each step in the assessment procedure. It was needed because there was an insufficient number of items in the pool to ensure that a desired item type would be available at every point in the procedure. Ideally, there would have been only one item type in each row of the table. In this case, all examinees would take the same sequence of item types. As can be seen in Table 2.ib, only an approximation to the ideal case was possible. Controlling the sequence of item types was intended to enhance the comparability of the latent trait (unidimensionality) across examinees.

The item-selection algorithm employed in the study had three phases: (1) selection of initial item, (2) up-down procedure, (3) maximum-likelihood procedure.

As Lord's (1977a) simulations had suggested the standard error of the BRTT would be relatively insensitive to the choice of the initial item, the same initial item was administered to each student. The item was selected from the first row and middle column of the table. For Form A, $b_1 = 1.38$; for Form B, $b_1 = 1.47$.

The maximum-likelihood estimation requires that the response vector contain at least one correct and one incorrect response. Following the initial item, a simple up-down rule was employed, raising or lowering the difficulty level of successive items until the complementary response was obtained. The current study employed a step size of ± one column.

Once the up-down procedure resulted in at least one item correct and one incorrect, the maximum-likelihood estimation procedure was used. The MLE algorithm operated as follows:

1. Determine the type of item to be administered by consulting the next row of the item-type table and selecting the column in which

$$b_i < \hat{\theta} \text{ and,}$$
$$|b_i - \hat{\theta}| \text{ was smallest}$$

2. Select the most discriminating item of the appropriate type and difficulty remaining in the pool and administer it.

2.5 Implementation of the Broad-Range Tailored Test

The Broad-Range Tailored Test was implemented on a PDP-11/40 computer system. A technical description of the system is provided in Swanson and Stocking (1977). This section presents an overview of the system's structure.

The following goals were established for the system design:

1. The system was to be designed in a flexible modular fashion to permit alteration of item pools, selection strategies, stopping rules, human-computer interaction protocols, and data collection strategies with minimal effort. The purpose of this objective was to facilitate the system's use in a variety of environments.

2. The system was to be coded in ANSI FORTRAN with minimal dependency on charscteristics of the PDP-11/40 computer. The purpose of this objective wss to fscilitace transportability of the softwsre to other computers.

3. The system was to be as independent of the Broad-Range Tailored Test as possible. The design of the system should permit most parameters to be specified at run time.

4. The human-computer interaction protocols were to be simple and natursl. The student should not perceive computerized administration as a barrier to overcome.

Figure 2.5a shows the file structure of the CAT system. The system employs five files. The first file is the item pool. Each item stored in the pool is assigned a number. The system builds sn index to the pool. The index contains information about each item including: the key, the item parsmeters (a, b, c), and the item type. Also contsined on the item pool file are the instructions which explsin how the item is to be answered. The system provides for two levels of instructions for each item type. The first level is s long form, the second level is a terse form. If specified, the system will present the long form instructions when the testee first encounters the item type and will present the second form on subsequent presentstions of the same type. This reduces the testee's resding losd.

The second file is called the test specification file. This file contains information which directs the system as to how the test is to be presented. Among the data in this file are: the item pool to be used, the item selection strategy to be employed, the scoring method to be employed, number of items to be administered, and feedback and re-try specifications. The file contains multiple test specifications, the choice of which is used is made when the test is actually administered.

The third file is called the message file. It contains messages to be displayed on the terminal if errors arise.

The fourth file is called the instructional file. It contains instructional frames that teach the student how to use the system. This feature was not used in the present study.

The fifth file is called the log file. The system writes a record to this file following each item. The file contains such information as: the testee identification, the item administered, the response, the response latency, and the current ability estimate. This file was used for data collection purposes in the current study. The system can be instructed to write log files at several levels of detail.

The items are displayed on a terminal connected to the computer via telephone lines. The terminal employed was a DEC VT52 cathode-ray tube display terminal. This terminal has a screen which displays 24 lines of 80 characters. It has a full typewriter keyboard and a small keypad containing 19 keys.

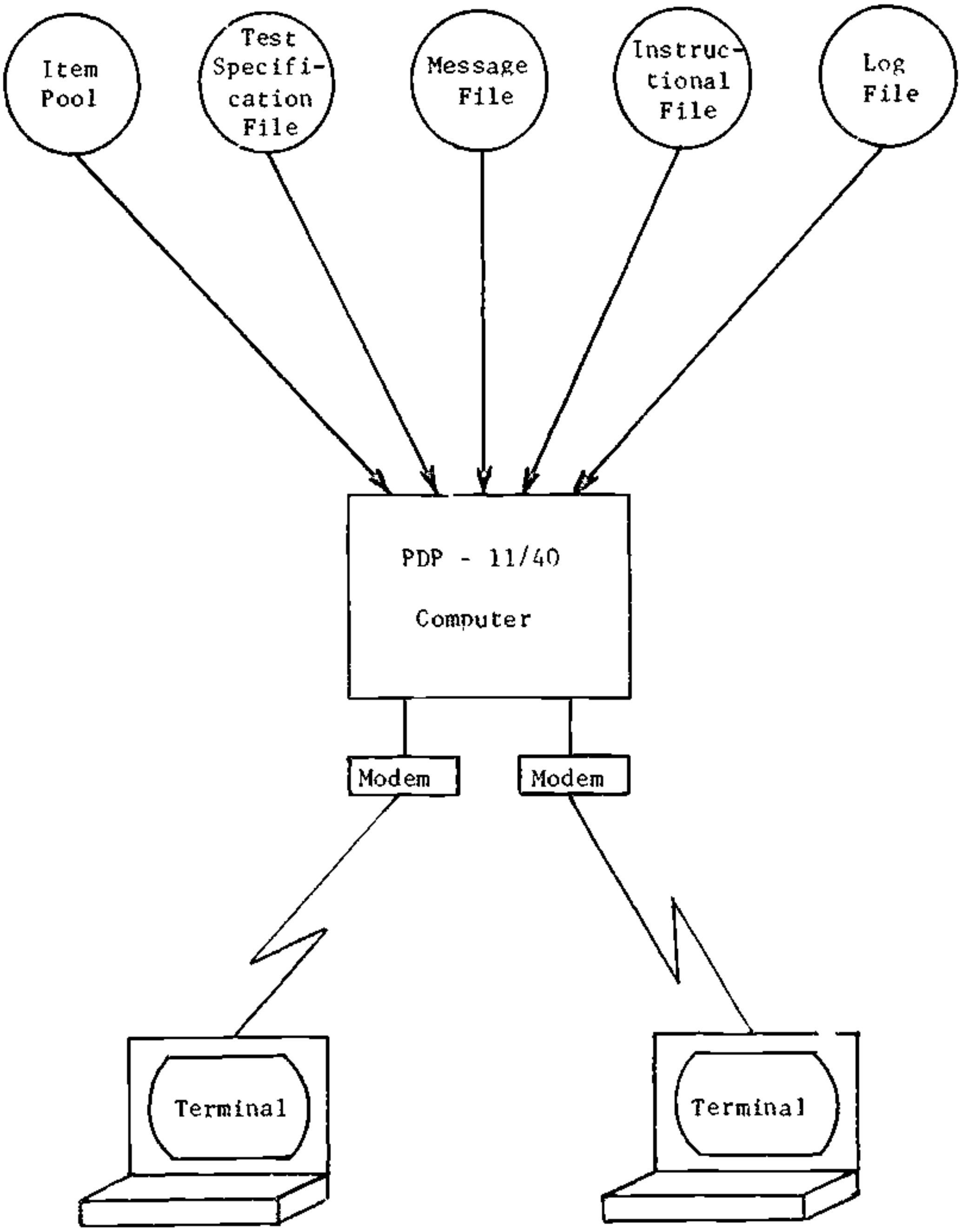


Figure 2.5a File Structure of the CAT System

To simplify the human-computer interaction, special keycaps were ordered for the small keypad. These keycaps are shown in Figure 2.5b. The student indicated his or her response by pressing the appropriate key. An asterisk appeared next to the corresponding option on the screen. The student could alter the response by pressing a different key or go on to the next item by pressing the key marked "enter." The "retrans" key instructed the computer to retransmit the item in case noise in the telephone system produced a garbled display.

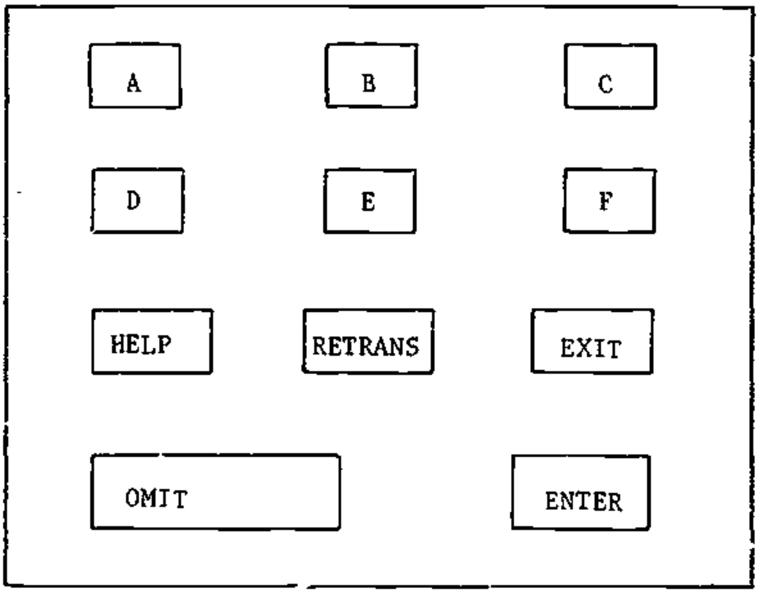


Figure 2.5b Response Keypad Used in the Study

Chapter III

Empirical Performance of the Broaa Range Tailored Test

3.1 Subjec and Method

In his article reporting on results of the simulation studies of the Broad Range Tailored Test (BRTT), Lord (1977) recommended the administration f the test to a live population -- a suggestion echoed by McBride (1976). This study was designed to explore the empirical performance of the BRTT. It involved two major tasks: (1) the development of a computer system capable of administering two forms of the BRTT and (2) the administration of the BRTT to a population of students. This study, which describes the experiment, is supplemented by the description of the computer system reported by Swanson and Stocking (1977).

Population

The range of individuals for whom the BRTT is appropriate is quite wide; it yields a score from the fifth grade to graduate level. However, a more homogeneous population was selected for this study, so that we could compare the performance of the BRTT with a conventional test of more limited range. Since prior simulations of the BRTT involved comparisons with the PSAT, the present study employed a comparable high school population. The inability of the PDP-11 computer system to support more than two "dial-up" terminals at a time significantly limited the number of individuals to whom the test could be administered. Therefore, an unselected population of high school students was used, drawing as many as possible from the eleventh grade. Each student received two forms of the BRTT; PSAT scores were obtained for those students who had taken the test as part of their academic work. Our target sample size was 150 students.

Selection of Schools

Frequency distributions of all high schools in Monmouth, Middlesex, Mercer, Hunterdon, and Somerset counties (NJ) were obtained for the number of students in the eleventh grade, the number of students who elected to take the PSAT, and the distribution of PSAT scores within school for the year 1977. This survey was conducted to select institutions with a reasonably wide and representative range of student abilities. Two of the six schools initially contacted agreed to participate in the experiment: Princeton High School and South Brunswick High School. Princeton High School requested that students not be paid for participation in the study, while South Brunswick requested that students be paid a fee of $3.50.

Sample Characteristics

One hundred forty-six students from the two schools participated in the study (Princeton N=80; South Brunswick N=66). Seventy-one of the students were males and seventy-five were females.

Experimental Design and Procedure

Each student received two forms of the BRTT, administered in a single class period. The order of the forms was counterbalanced within sex, as shown below:

	Form	
Male	A	B
Female	A	B
Male	B	A
Female	B	A

Assignment of students to order was performed randomly within sex.

Upon entering the testing room, the student was "signed on" the computer by a proctor who explained the use of the terminal. The student then responded to the 25 items selected by the computer. Following completion of the first form of the BRTT, the proctor initiated the alternate form and the student responded to an additional 25 items. The student was then asked to complete the posttest questionnaire.

3.2 Overview of Data Analyses

Conceptually, the data analyses presented in this chapter may be divided into four units. The first unit is descriptive and presents characteristics of the observed data. In particular, Section 3.3 describes the scores and score transformations derived from the data; Section 3.4 presents frequency distributions of the scores; and Section 3.5 analyzes the distributional characteristics of the scores to determine whether they meet the normality assumptions for correlational statistics.

The second unit, which presents information functions for both forms of the BRTT, is directly related to simulation data reported by Lord (1977a). The information functions are presented in Section 3.6.

The third unit of data analyses involves the reliability and validity of the BRTT. Section 3.7 shows the comparison of the reliability coefficients for the scores and transformations. The correlations between the BRTT scores and the PSAT verbal scores are presented as a measure of concurrent validity. A series of tests on the mean scores from Forms A and B are presented, which bear on the parallelism of the two forms. The reliabilities presented in Section 3.7 were computed at the final (25th) item. The reliabilities of the BRTT with the PSAT at test lengths of 1, 2,...., 25 items are compared in Section 3.8, and are presented in the form of plots. Discussion on the likelihood function and the maximum likelihood estimator is presented in Section 3.9.

The fourth unit of data analyses involves the performance of the maximum-likelihood estimator (MLE). Section 3.10 presents analyses of the MLE and demonstrates that, overall, the procedure performed as expected. Section 3.11 is a Monte Carlo analysis of the MLE procedure, in which the examinee's observed responses compared with responses obtained by simulation based on the estimated Θ. Section 3.12 presents the examinee response patterns which resulted in anomalous ability estimates.

The findings presented in this chapter are summarized in Section 3.13.

3.3 Observed Scores and Score Transformations

The characteristics of nine score variables derived from the data on the computer log files were examined. The scores are defined as follows:

1. Θ. Theta is the parameter ccmmonlv used to denote the latent trait. Each examinee is assumed to possess a fixed value for Θ, which is estimated from his or her responses to test items.

2. Ω. Omega is a monotone transformation of Θ which was proposed by Lord (1975). Lord proposed this transformation because of an observed correlation among item parameters de ined on the Θ scale. The transformation is:

$$\Omega\ (\Theta) = \int_{-\infty}^{\Theta} R(a|b)\,db$$

Where $R(a|b)$ is the observed regression of a, the discrimination onto b, the difficulty. The transformation eliminates correlation between the a and b parameters. The regression was taken to be linear in the variable b and the res ing transformation $\Omega(\Theta)$ is an eighth degree polynomial in Θ with coefficients (starting with the constant, ending with the eighth power)·

j	Coefficient of Θ^j
0	0.01916364510799269000
1	0.52781849849492630000
2	0.02985677403820969000
3	-0.00140773413450821900
4	-0.00039316355724225770
5	-0.00003302234217826338
6	-0.00000142120205848321
7	-0.00000003151838404393
8	-0.00000000028208952766

3. Number right. The common score employed in paper-and-pencil tests. Other scores related to number right that are occasionally referenced are number wrong and number omitted.

4. p. Expected proportion right is the true score transformation for the test (Lord and Novick, 1969, p. 387), and is computed over the entire item pool by the formula:

$$p = \frac{1}{N} \Sigma_{i=1,N} P_i(\theta)$$

where N is equal to the number of items in the pool, and $P_i(\theta)$ is the probability of correct response to the ith item at ability θ.

5. b_{total} - is the mean difficulty over all items administered.

6. $b_{correct}$ - is the mean difficulty over all items answered correctly.

7. b_{high} - is the highest difficulty of all items answered correctly.

8. b_{final} - is the difficulty of the final item administered.

9. S,T - are a weighted number correct score and an expected weighted number correct score. They are given by the formulas:

$$S(\theta) = \Sigma_{i=1,25}\, u_i\, w_i(\theta)$$

and $$T(\theta) = \Sigma_{i=1,25}\, P_i(\theta)\, w_i(\theta)$$

where $$w_i(\theta) = \frac{\dot{P}_i(\theta)}{P_i(\theta)\, Q_i(\theta)}$$

and $u_i = 1$ if the individual responded correctly to the item, 0 otherwise.

These scores serve as a final check on the adequacy of the algorithm for estimating θ.

3.4 Frequency Distribution of Scores

Frequency distributions for Θ, Ω, number correct, number wrong, number omitted, average difficulty of items answered correctly, and expected proportion correct over the entire item pool are presented in this section.

Eleven cases have been deleted from the present and subsequent analyses. In Form A, six individuals obtained Θ scores of -6.00. These individuals were excluded from the analyses because they represented anomalous cases whose scores were not directly comparable to the remainder of the subjects. Five additional cases were excluded after finer analysis revealed several anomalies either in the individuals' responses or in the computation of Θ by the computer system. (See Section 3.12 for a discussion of the anomalous protocols).

Table 3.4a summarizes the frequency distribution for Θ computed at the final item for Forms A and B of the BRTT. In both cases, the majority of the examinees scored in the range -0.5 to 2.5. Table 3.4b summarizes the frequency distribution of Ω for Forms A and B. Table 3.4c summarizes the number of items answered correctly for Forms A and B. For both Θ and Ω, the Form A and Form B distributions appear roughly comparable.

Table 3.4d summarizes the distribution of the number of items answered incorrectly for both forms of the BRTT; Table 3.4e summarizes the distribution of the number of items omitted for the two forms. Note that for both forms, at least 50% of all examinees omitted one or more items.

Additional techniques for scoring omitted items and allowing examinees to review should be important areas of research in the future, si-ce some examinees have a high tendency to omit items.

Table 3.4f summarizes the distribution of the average difficulty of items answered correctly; Table 3.4g summarizes the distribution of the expected proportion of items answered correctly if the examinee answered all the items in the pool. The Form B distribution of each score is slightly more dispersed than the Form A distribution of the corresponding score. Section 3.7 and 3.8 provide information which bears on the comparability of the two forms. Table 3.7c presents the frequency distributions of the PSAT scores.

TABLE 3.4a

Frequency Distribution of θ

	Form A		Form B	
Interval	Freq.	%[†]	Freq.	%
4.0 - 3.3	1	0.7	1	0.7
3.3 - 2.6	4	3.0	1	0.7
2.6 - 1.9	20	14.8	26	19.3
1.9 - 1.2	42	31.1	30	22.2
1.2 - 0.5	30	22.2	38	28.1
0.5 - -0.2	26	19.3	25	18.5
-0.2 - -0.9	9	6.7	11	8.1
-0.9 - -1.6	2	1.5	2	1.5
-1.6 - -2.3	0	0.0	1	0.7
-2.3 - -3.0	1	0.7	0	0.0
N	135		135	
Mean	1.0702		0.9829	
SD	0.9598		0.9550	
Minimum Value	-2.6118		-1.7689	
Maximum Value	3.7837		3.3755	

[†] Total of percentages throughout is not 100 due to rounding.

TABLE 3.4b

Frequency Distribution of Ω

	Form A		Form B	
Interval	Freq.	%	Freq.	%
2.5 - 2.1	1	0.7	0	0.0
2.1 - 1.7	2	1.5	2	1.5
1.7 - 1.3	10	7.4	13	9.6
1.3 - 0.9	25	18.5	26	19.3
0.9 - 0.5	50	37.0	32	23.7
0.5 - 0.1	26	19.3	32	23.7
0.1 - -0.3	16	11.9	24	17.8
-0.3 - -0.7	4	3.0	4	3.0
-0.7 - -1.1	0	0.0	2	1.5
-1.1 - -1.5	1	0.7	0	0.0

	Form A	Form B
N	135	135
Mean	0.6359	0.5852
SD	0.5417	0.5442
Minimum Value	-1.1453	-0.8165
Maximum Value	2.2567	2.0191

TABLE 3.4c

Frequency Distribution of Number of Items Correct

for Forms A and B

Number	Form A		Form B	
Correct	Freq.	%	Freq.	%
22	1	0.7	0	0.0
21	1	0.7	3	2.2
20	1	0.7	2	1.5
19	10	7.4	3	2.2
18	8	5.9	15	11.1
17	20	14.8	19	14.1
16	17	12.6	18	13.3
15	18	13.3	22	16.3
14	27	20.0	23	17.0
13	13	9.6	14	10.4
12	13	9.6	8	5.9
11	3	2.2	6	4.4
10	1	0.7	1	0.7
9	2	1.5	1	0.7
N	135		135	
Mean	15.1333		15 2296	
SD	2.4089		2.3434	

TABLE 3.4d

Frequency Distribution of Number of Items Incorrect

for Forms A and B (N = 135)

Number	Form A		Form B	
Incorrect	Freq.	%	Freq.	%
16	1	0.7	0	0.0
15	1	0.7	0	0.0
14	1	0.7	3	2.2
13	6	4.4	2	1.5
12	9	6.7	8	5.9
11	12	8.9	12	8.9
10	22	16.3	22	16.3
9	15	11.1	21	15.6
8	18	13.3	25	18.5
7	18	13.3	22	16.3
6	13	9.6	8	5.9
5	11	8.1	7	5.2
4	4	3.0	2	1.5
3	3	2.2	3	2.2
2	1	0.7	0	0.0
N	135		135	
Mean	8.5037		8.5852	
SD	2.6818		2.2540	

TABLE 3.4e

Frequency Distribution of Number of Items Omitted

for Forms A and B (N=135)

Number	Form A		Form B	
Omitted	Freq.	%	Freq.	%
7	2	1.5	2	1.5
6	2	1.5	1	0.7
5	5	3.7	4	3.0
4	6	4.4	5	3.7
3	16	11.9	13	9.6
2	18	13.3	16	11.9
1	25	18.5	29	21.5
0	61	45.2	65	48.1
N	135		135	
Mean	1.3630		1.1852	
SD	1.6867		1.5797	

TABLE 3.4f

Frequency Distribution of Average Difficulty for

All Items Answered Correctly

	Form A		Form B	
Interval	Freq.	%	Freq.	%
2.5 - 3.5	1	0.7	0	0.0
1.5 - 2.5	28	20.7	32	23.7
0.5 - 1.5	66	48.9	48	35.6
-0.5 - 0.5	31	23.0	43	31.9
-1.5 - -0.5	7	5.2	10	7.4
-2.5 - -1.5	1	0.7	2	1.5
-3.5 - -2.5	1	0.7	0	0.0

	Form A	Form B
N	135	135
Mean	0.8055	0.7213
SD	0.8536	0.8732
Minimum Value	-3.034[illegible]	-1.8044
Maximum Value	2.6471	2.4312

TABLE 3.4g

Frequency Distribution of Expected Proportion Correct for Entire Item Pool

	Form A		Form B	
Interval	Freq.	%	Freq.	%
1.0 - 0.9	1	0.7	2	1.5
0.9 - 0.8	14	10.4	23	17.0
0.8 - 0.7	40	30.0	29	21.5
0.7 - 0.6	40	30.0	43	31.9
0.6 - 0.5	27	20.0	28	30.7
0.5 - 0.4	10	7.4	8	5.9
0.4 - 0.3	2	1.4	2	1.5
0.3 - 0.2	1	0.7	0	0.0
0.2 - 0.1	0	0.0	0	0.0
0.1 - 0.0	0	0.0	0	0.0
N	135		135	
Mean	0.66244		0.66712	
SD	0.118111		0.11831	
Minimum Value	0.28988		0.35343	
Maximum Value	0.93656		0.92536	

3.5 Tests for Normality of Score Distributions

Inferences that use the usual correlational statistics to evaluate the reliability of the BRTT, are based on the assumption that score distributions are approximately normal. Data which deviate from this assumption may produce inflated correlations. The distributional characteristics of the data were investigated by plotting the percentile from the standard normal distribution (z) against the same percentile from the various empirically observed score distributions. Whenever the empirical values follow the normal distributions exactly, the points fall precisely on a straight line, and deviations from the straight line represent deviations from normality. Points which fall above the line represent observed values which exceed the expected values, whereas points which fall below the line represent values below expectation. Observations which tend to inflate correlations are those with percentile rank greater than 50 (less than 50) and with percentiles falling above (below) the line.

Figures 3.5a and 3.5b show the distribution of θ for Forms A and B respectively. Both plots show deviations from normality; under normality the Form A plot shows three negative values smaller than would be expected, and the Form B plot shows that the positive tail of the distribution is shorter than expected. These observations might suggest that a suitable transformation be made on the θ scores to achieve normality. However, since the resulting transformation would depend on this specific data set, it might not be suitable for other data sets. We determined not to transform the data because a reliability analysis based on the transformed data would not be of general use. Research to develop measures of reliability which are not overly dependent on the distribution of ability in the population is recommended.

The Pearson correlation coefficient, which is based on the scores described, is used as a measure of reliability in the remaining analyses. The sample Pearson correlation coefficient can be used to estimate the population Pearson coefficient. Since the design for this study is a close approximation to simple random sampling of subjects, the sample correlation coefficient is an unbiased estimate of the population parameter.

Figures 3.5c through 3.5n show the distributional characteristics of the scores θ, b_{total}, $b_{correct}$, b_{final}, b_{high}, and $\underline{p}$; all the scores show deviations from normality. The score showing the least departure from normality is $b_{correct}$. To further investigate the distribution of $b_{correct}$, a bivariate plot of Form A versus Form B is included in Figure 3.5o. This plot exhibits several peculiarities which could lead to rejection of normality upon a finer analysis than was conducted here.

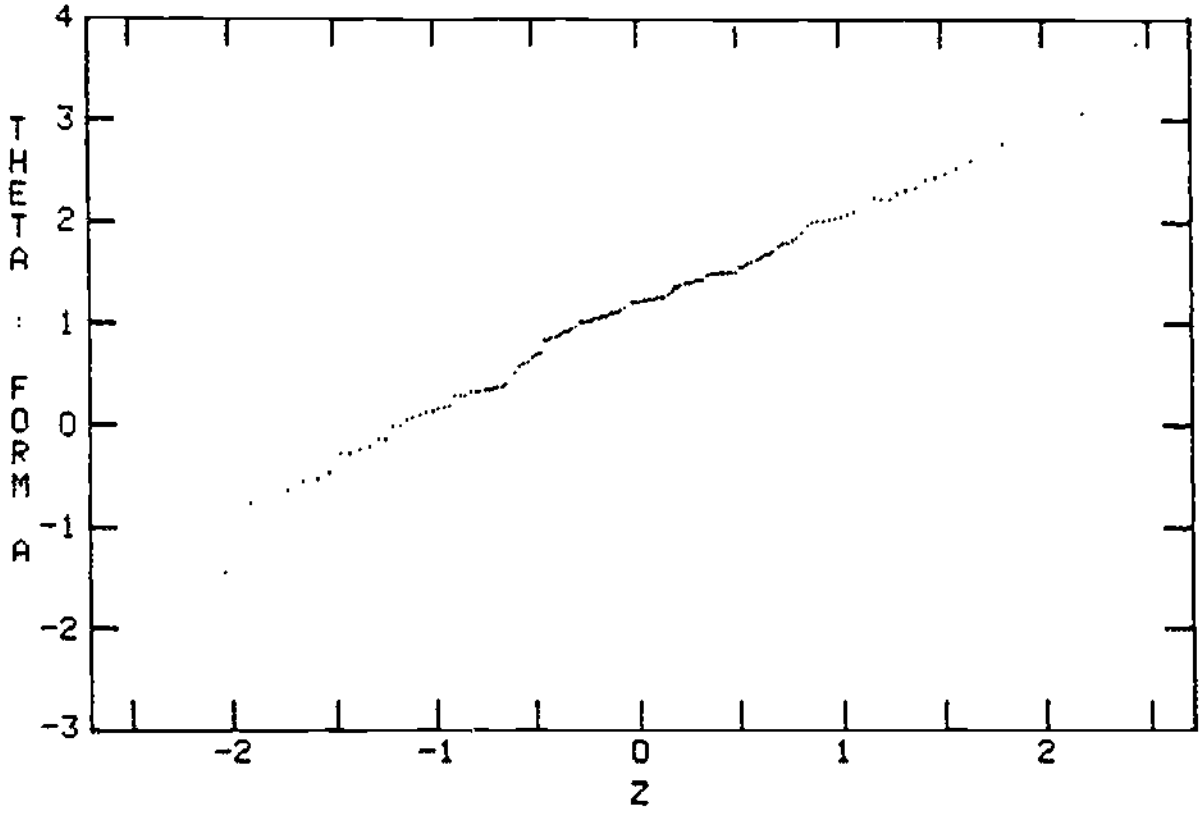


Figure 3.5a Distributional Characteristics of Theta (Form A)

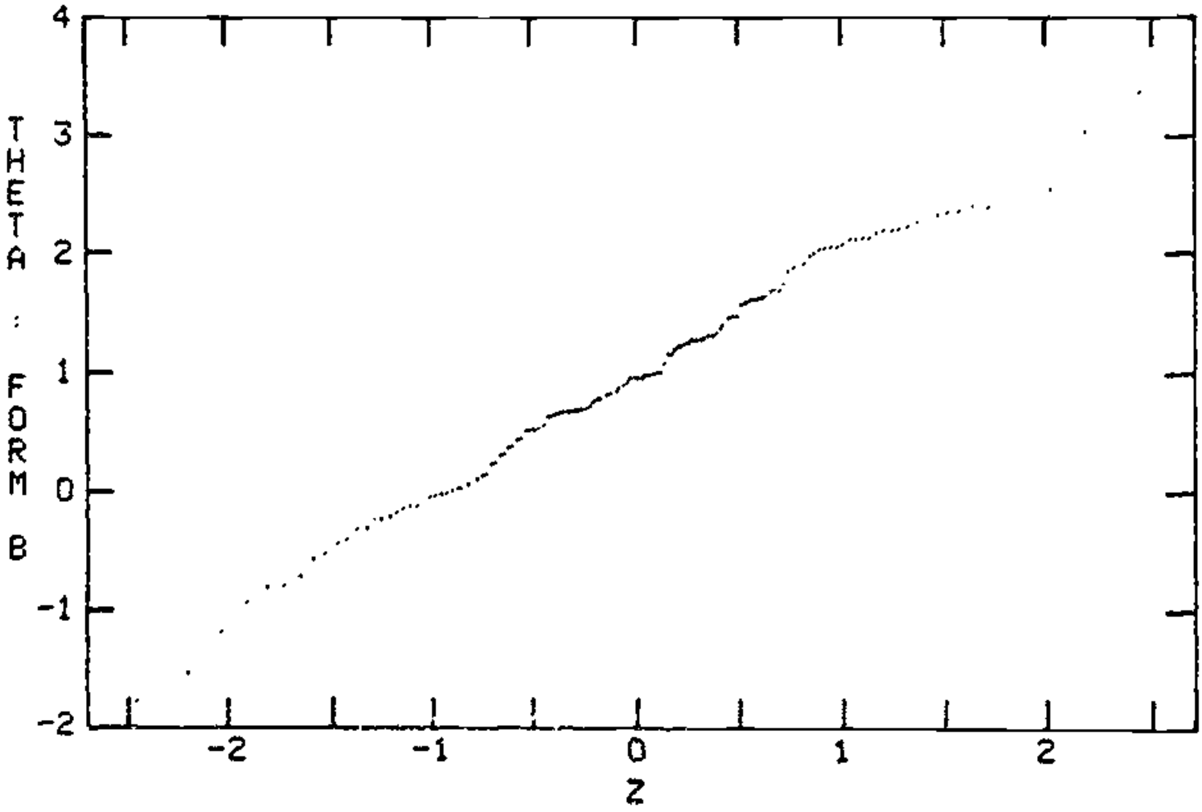


Figure 3.5b Distributional Characteristics of Theta (Form B)

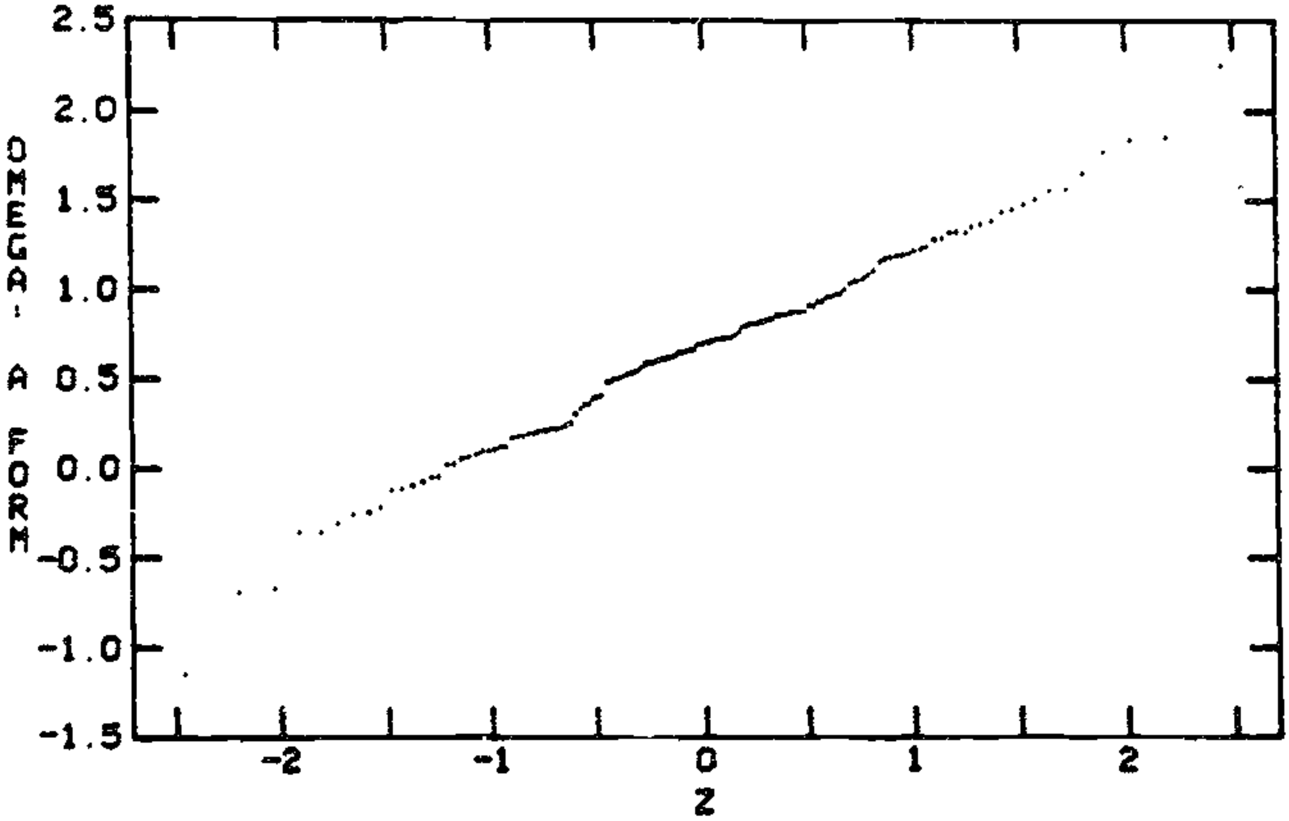


Figure 3.5c Distributional Characteristics of Omega (Form A)

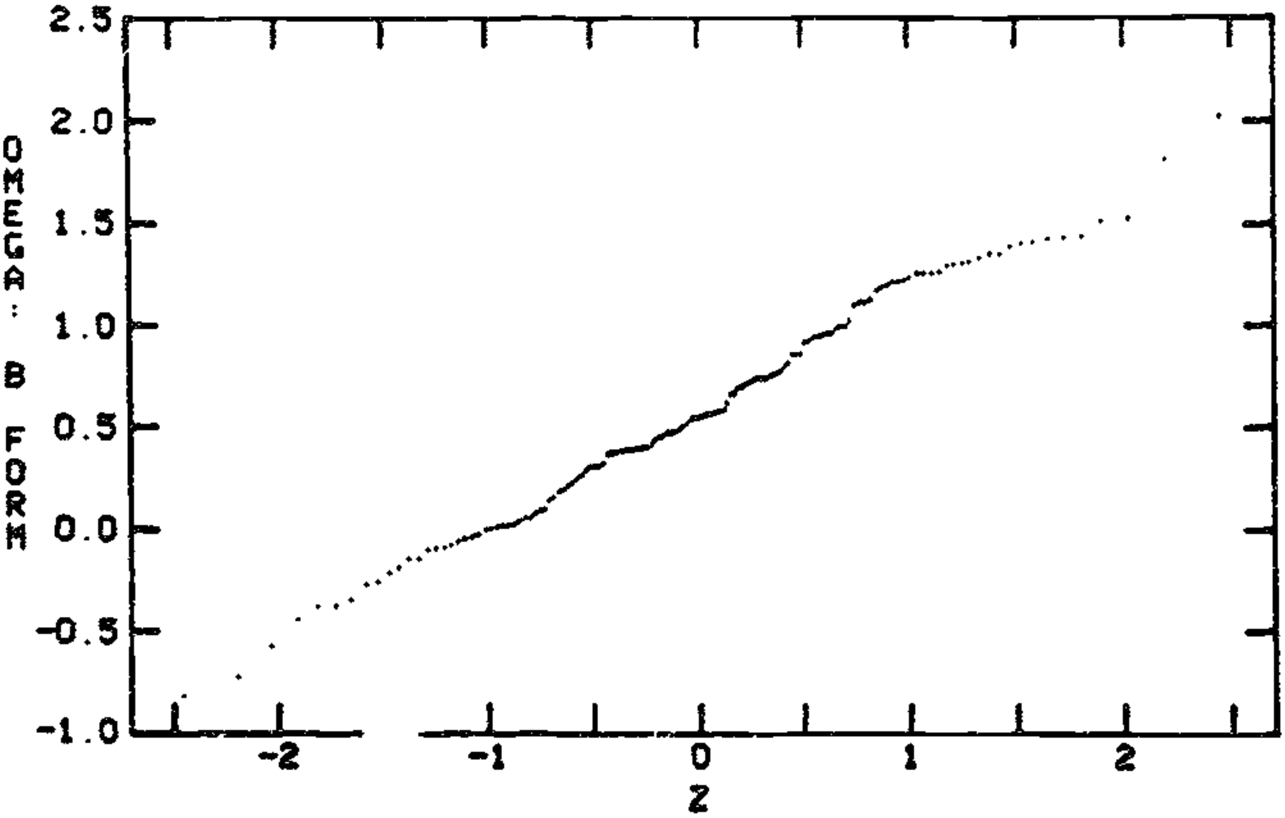


Figure 3.5d Distributional Characteristics of Omega (Form B)

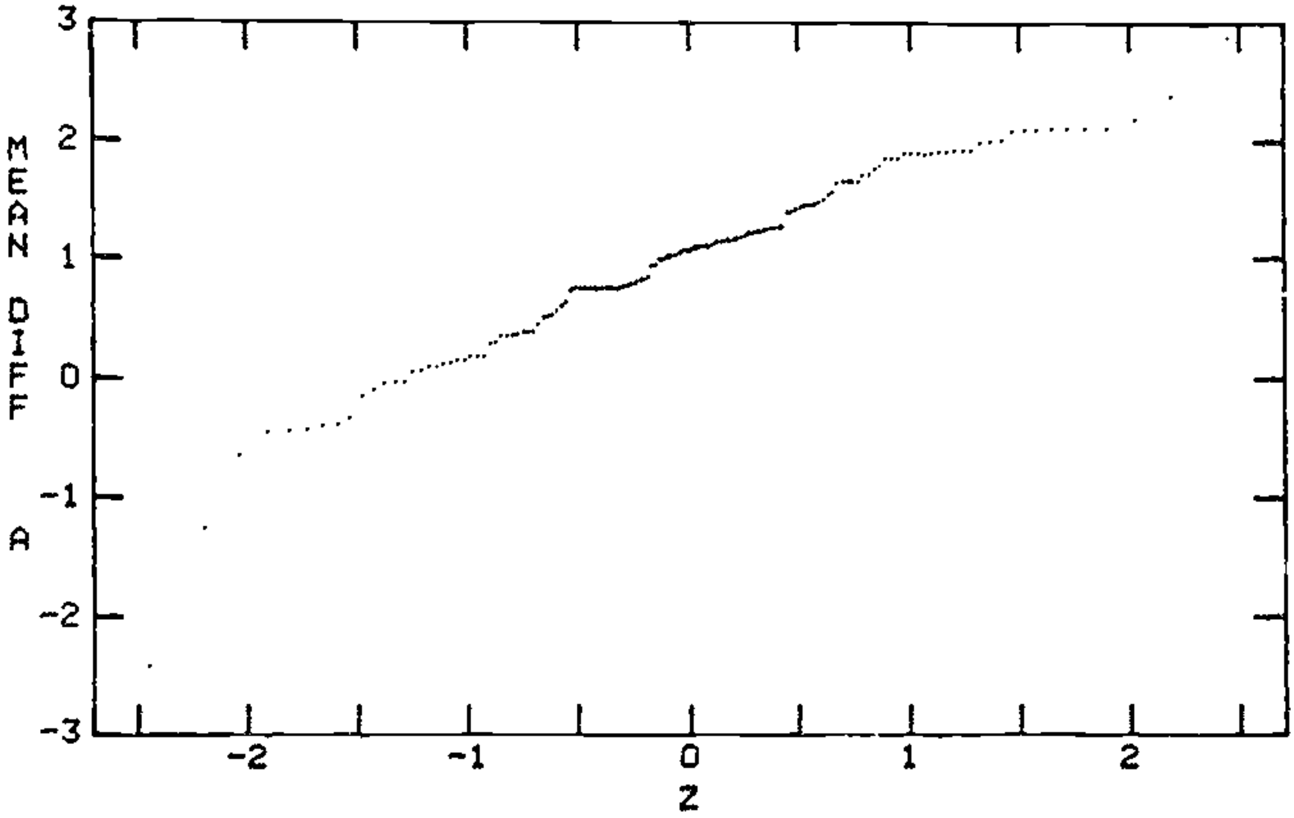


Figure 3.5e Distributional Characteristics of b_{total} (Form A)

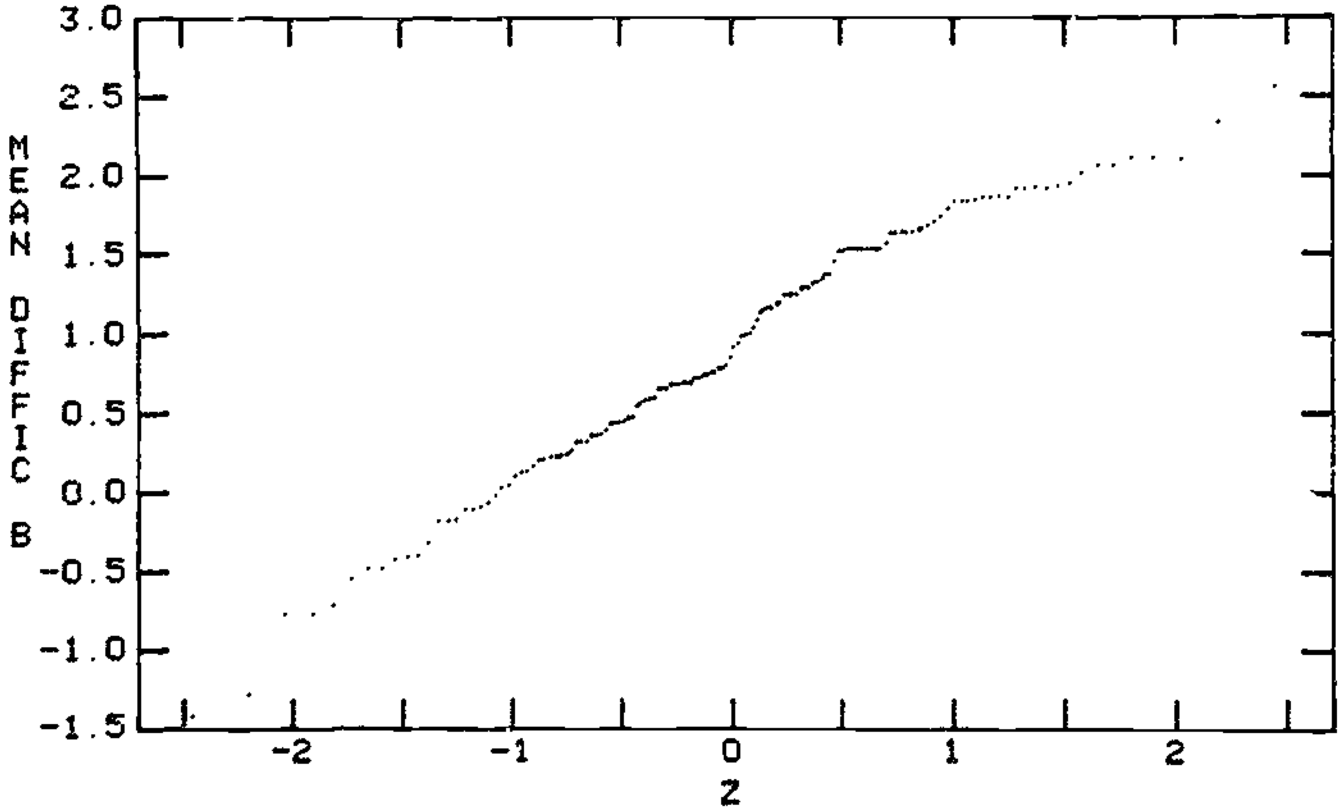


Figure 3.5f Distributional Characterics of b_{total} (Form B)

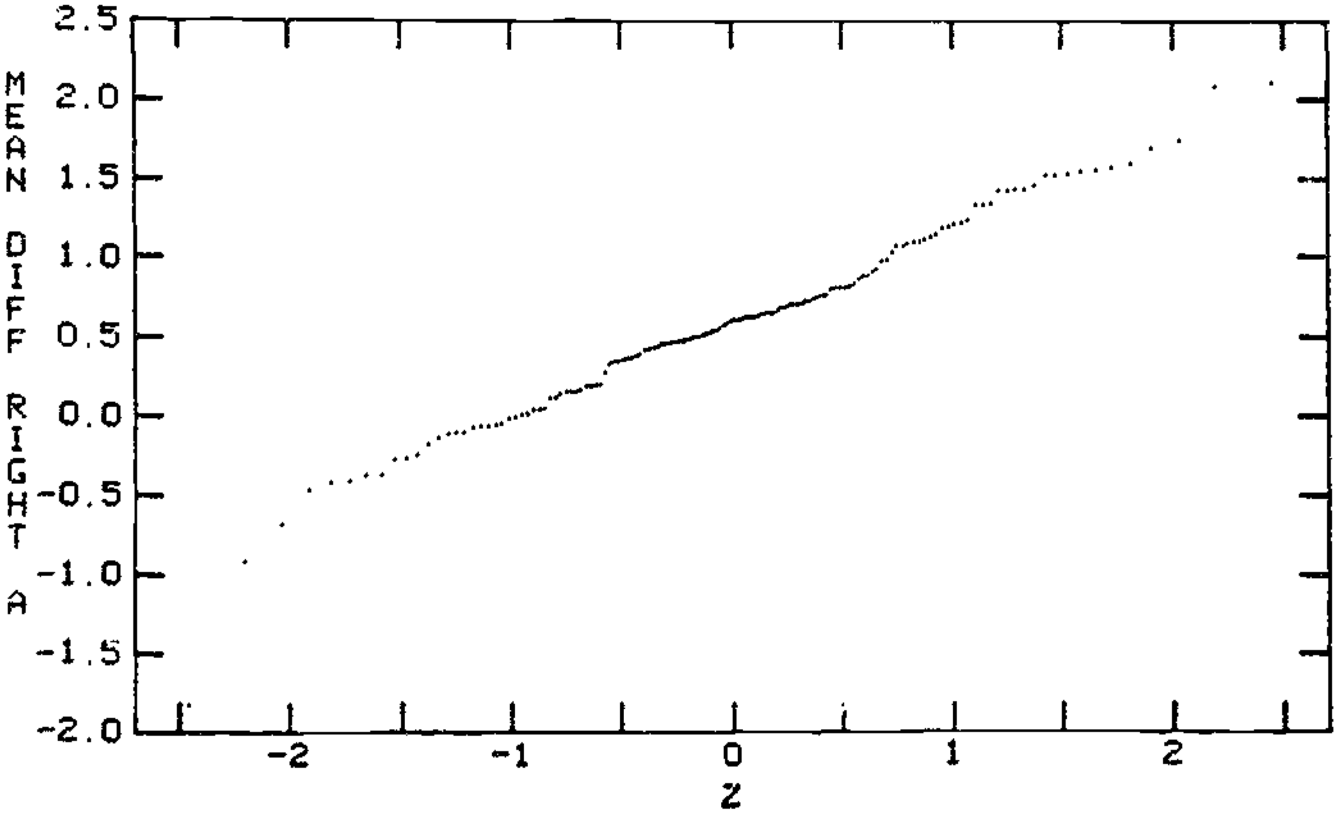


Figure 3.5g Distributional Characteristics of $b_{correct}$ (Form A)

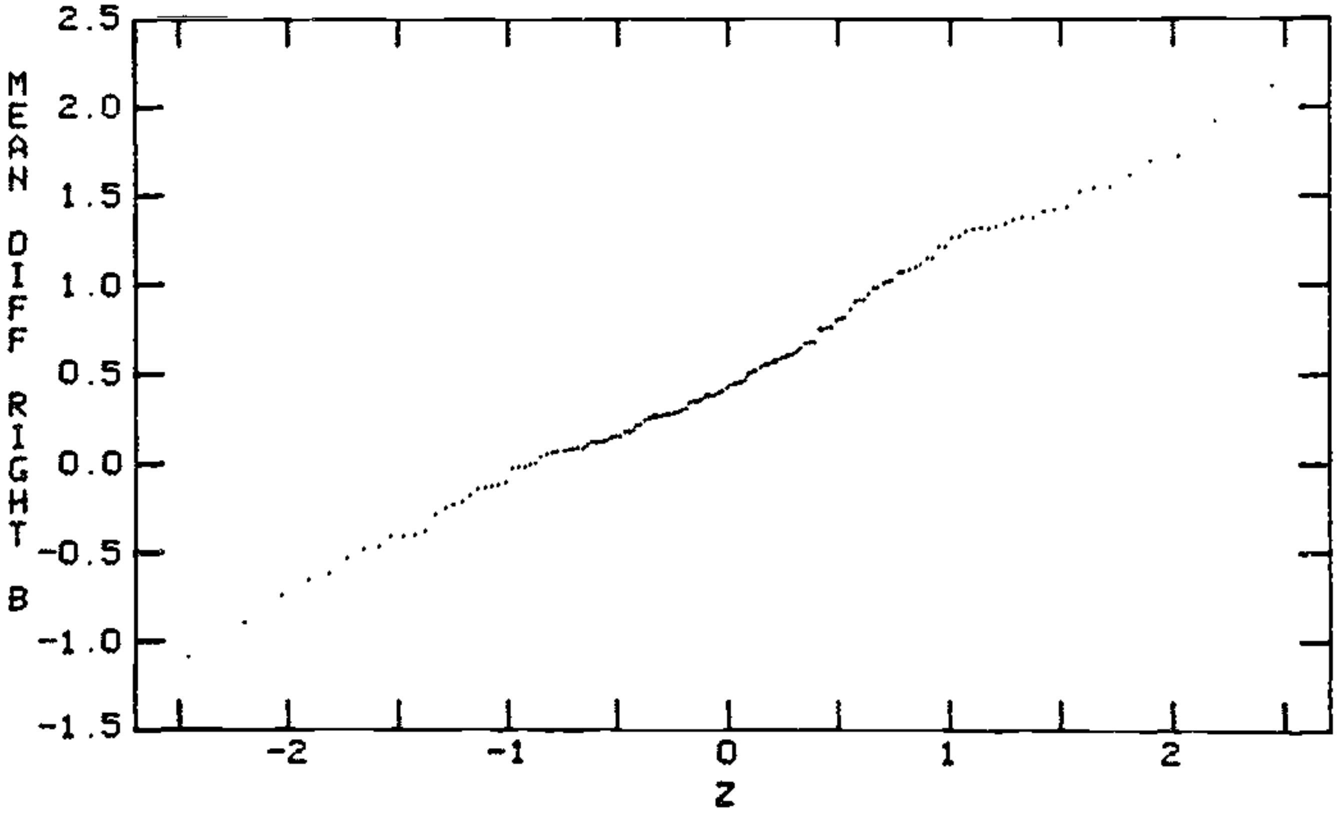


Figure 3.5h Distributional Characteristics of $b_{correct}$ (Form B)

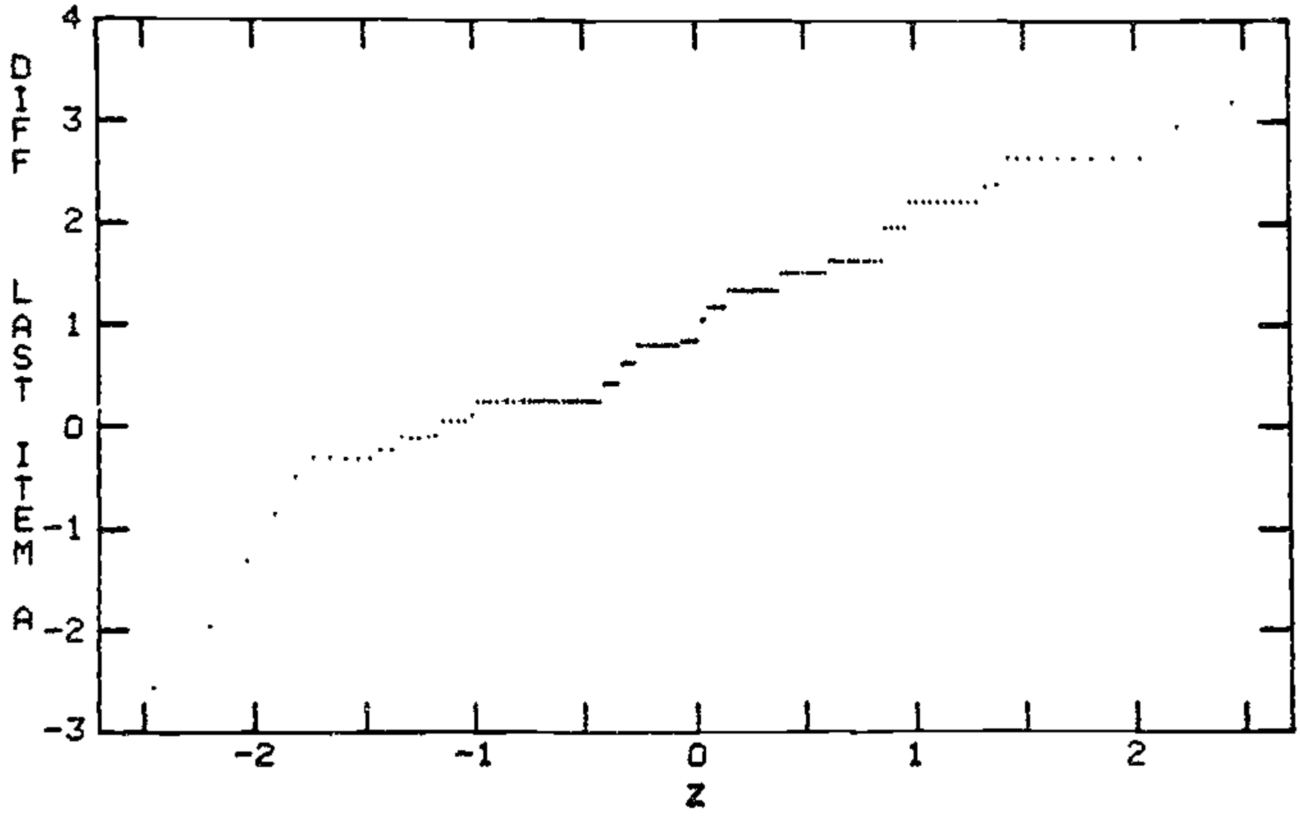


Figure 3.5i Distributional Characteristics of b_{final} (Form A)

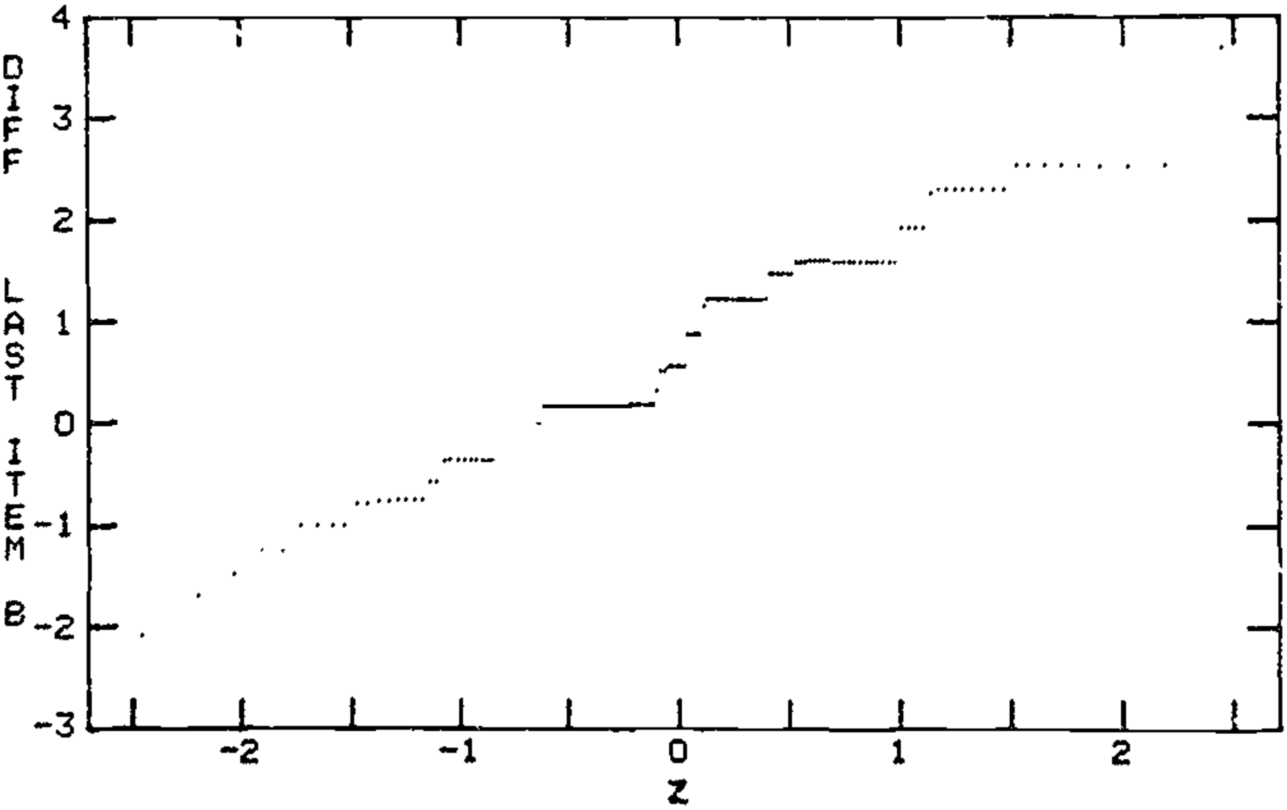


Figure 3.5j Distributional Characteristics of b_{final} (Form B)

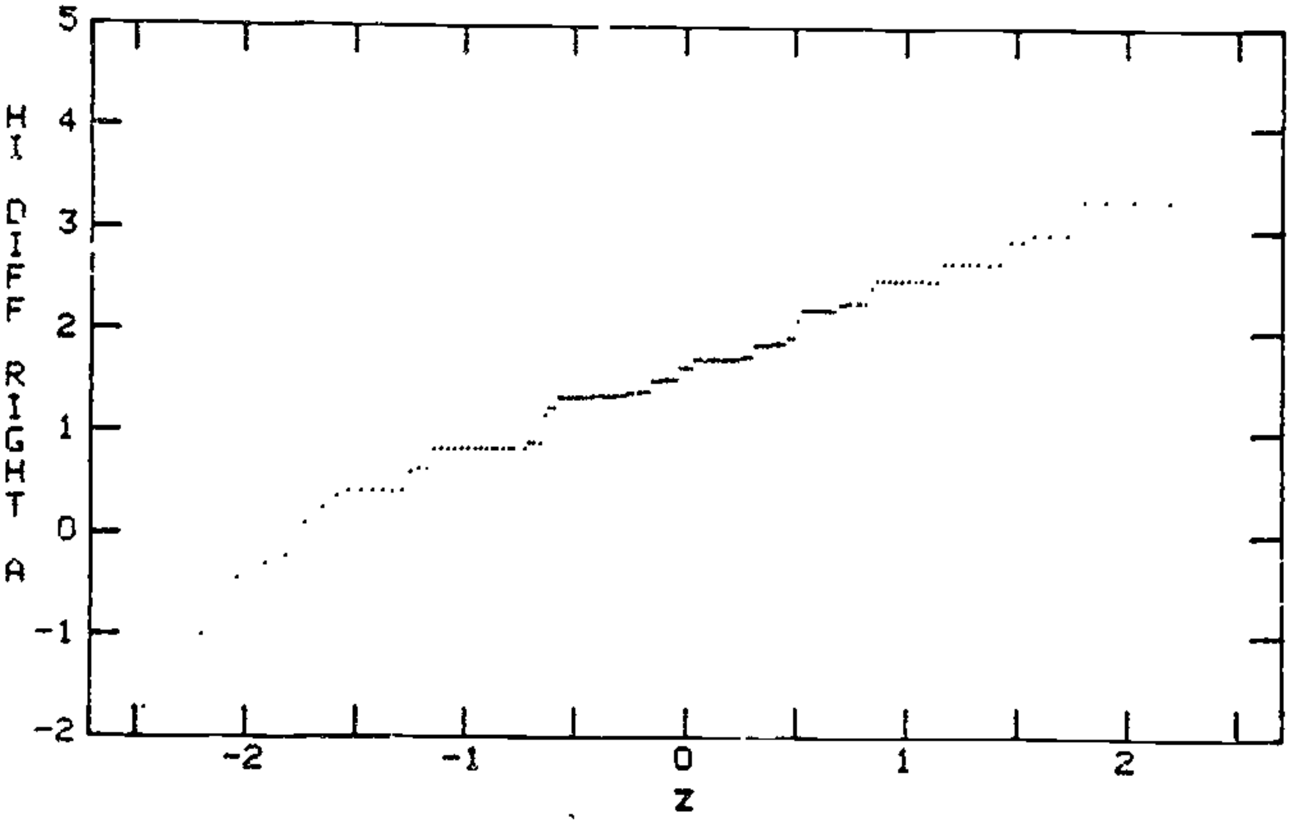


Figure 3.5k Di-tributional Characteristics of b_{high} (Form A)

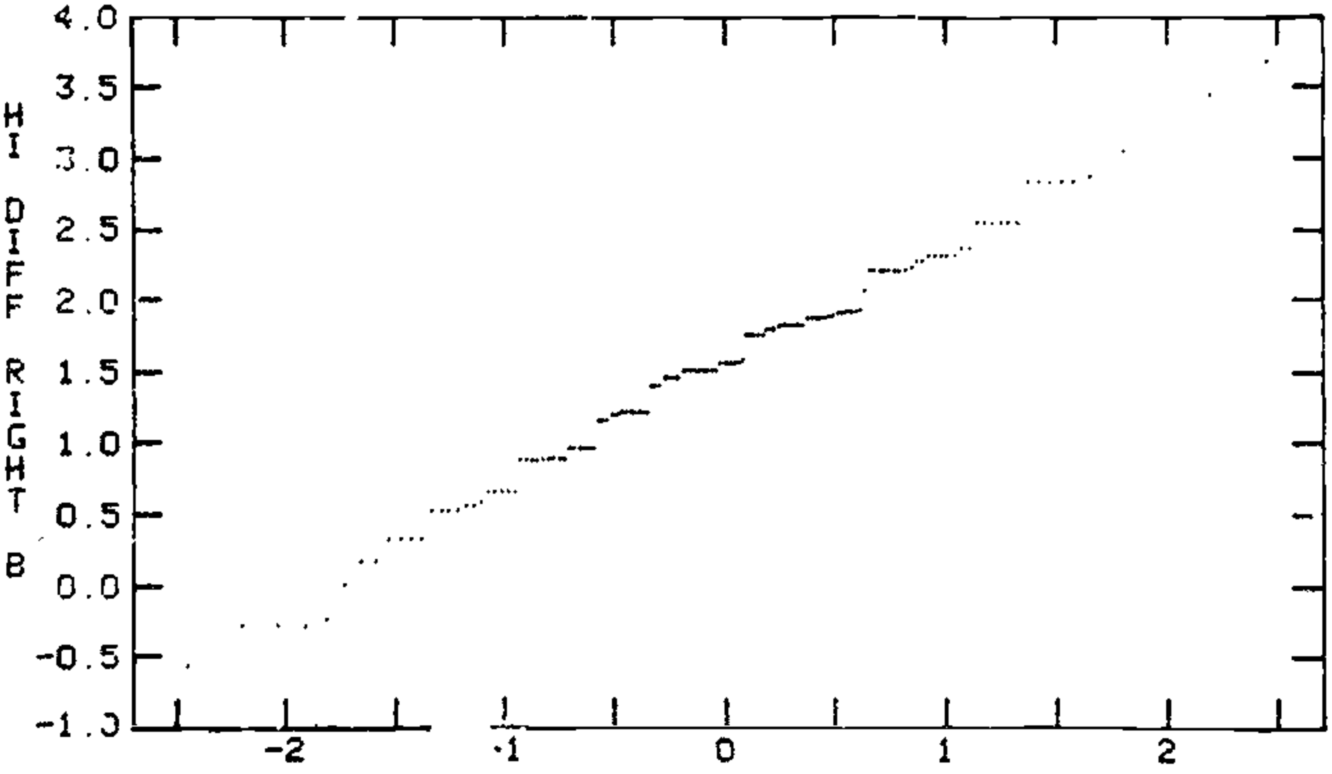


Figure 3.5L Distributional Characteristics of b_{high} (Form B)

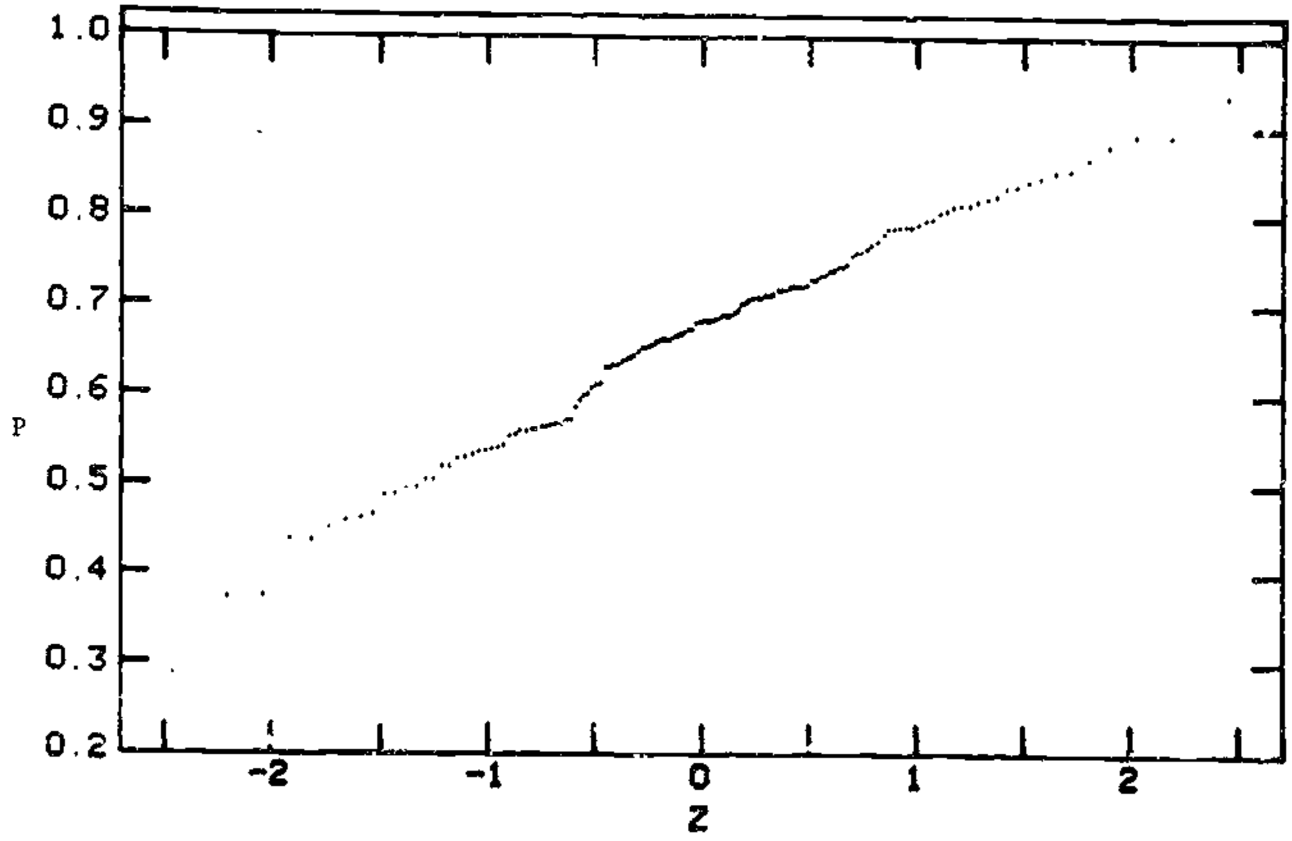


Figure 3.5m Distributional Characteristics of p (Form A)

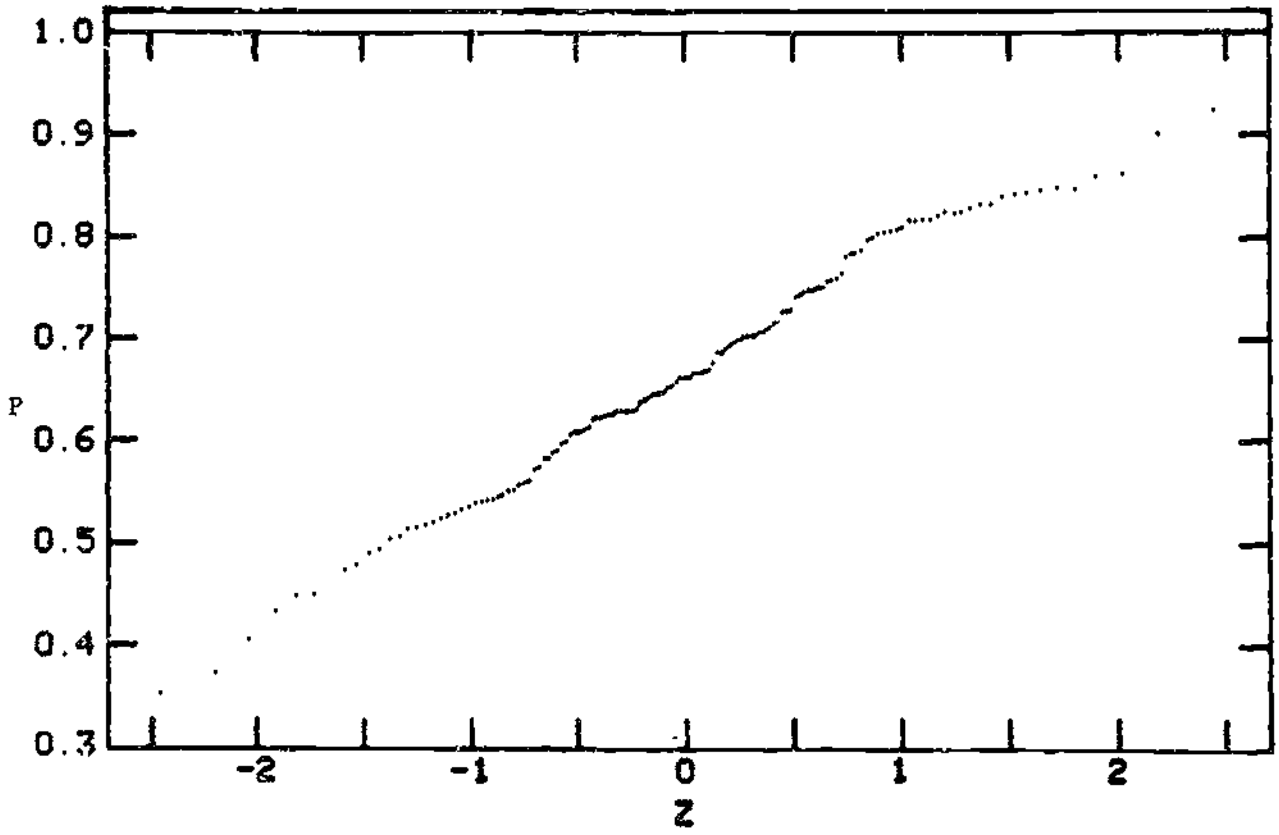


Figure 3.5n Distributional Characteristics of p (Form B)

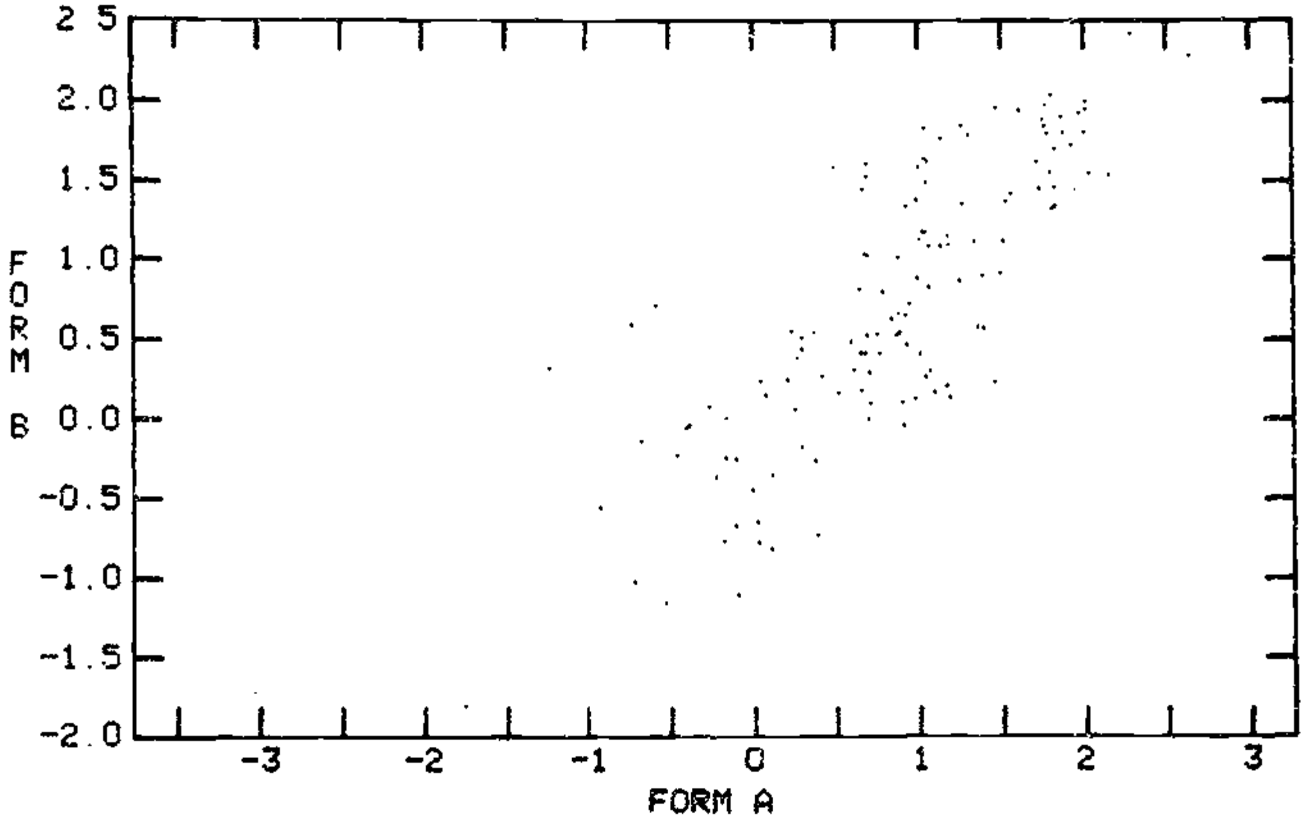


Figure 3.5o. Bivariate plot of $b_{correct}$.

3.6 Information Function

The information associated with the maximum likelihood estimate of the ability parameter θ is given by Lord and Novick (1967, p. 460) as:

$$I(\theta) = \sum_{i=1}^{n} [\dot{P}_i(\theta)]^2 / P_i(\theta)\ Q_i(\theta)$$

where $P_i(\theta)$ is the probability of a correct response to item i,
$Q_i(\theta) = 1 - P_i(\theta)$, and $\dot{P}_i(\theta)$ is the derivative of $P_i(\theta)$
with respect to θ.

For the BRTT, the value of n is 25 and the items may be different for different examinees.

The values of the information for each estimated ability level were computed based upon the actual items administered. The values were then transformed to obtain the information at each estimated Ω-score of ability. The information in Ω is

$$\tilde{I}(\Omega) = I(\theta)/[\dot{\Omega}(\theta)]^2$$

where $I(\theta)$ is given above, θ is given implicitly as $\Omega(\theta) = \Omega$ and $\dot{\Omega}(\theta)$ is the derivative of the transform $\Omega(\theta)$ with respect to θ.

Figures 3.6a and 3.6b display the scatter-plots of $\tilde{I}(\Omega)$ vs Ω for Forms A and B respectively. Each scatter-plot has been smoothed using a cubic spline interpolation available in the SPEAKEASY computing package. These results are displayed by the dashed line. In addition, the solid line in Figure 3.6a represents the simulated reciprocals of the variance of the maximum likelihood estimator of Ω given in Lord (1977a)

Figure 3.6a presents one of the most interesting results of this study: the simulated curve is a very close approximation to the actual outcome of a live experiment. The theory provides a useful tool in the evaluation of mental tests, but one should be cautioned that the simulated curve was obtained by simulating responses to items according to the item response model. Further, the empirical information function is calculated according to the theoretical model for responses based on those items chosen by the BRTT in the experiment.

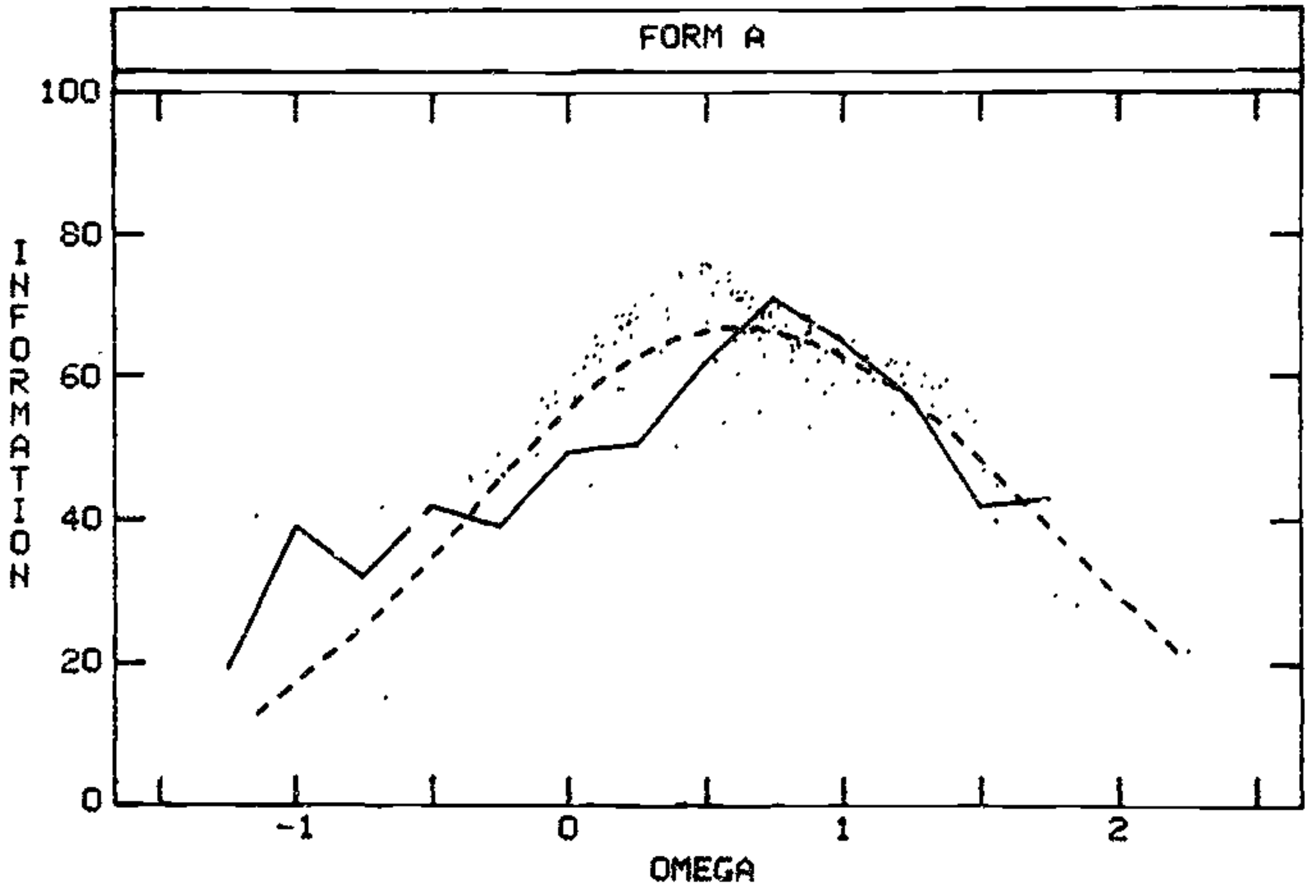


Figure 3.6a Observed information (dots), Smoothed information (dashes), Simulated information (solid line), (Form A).

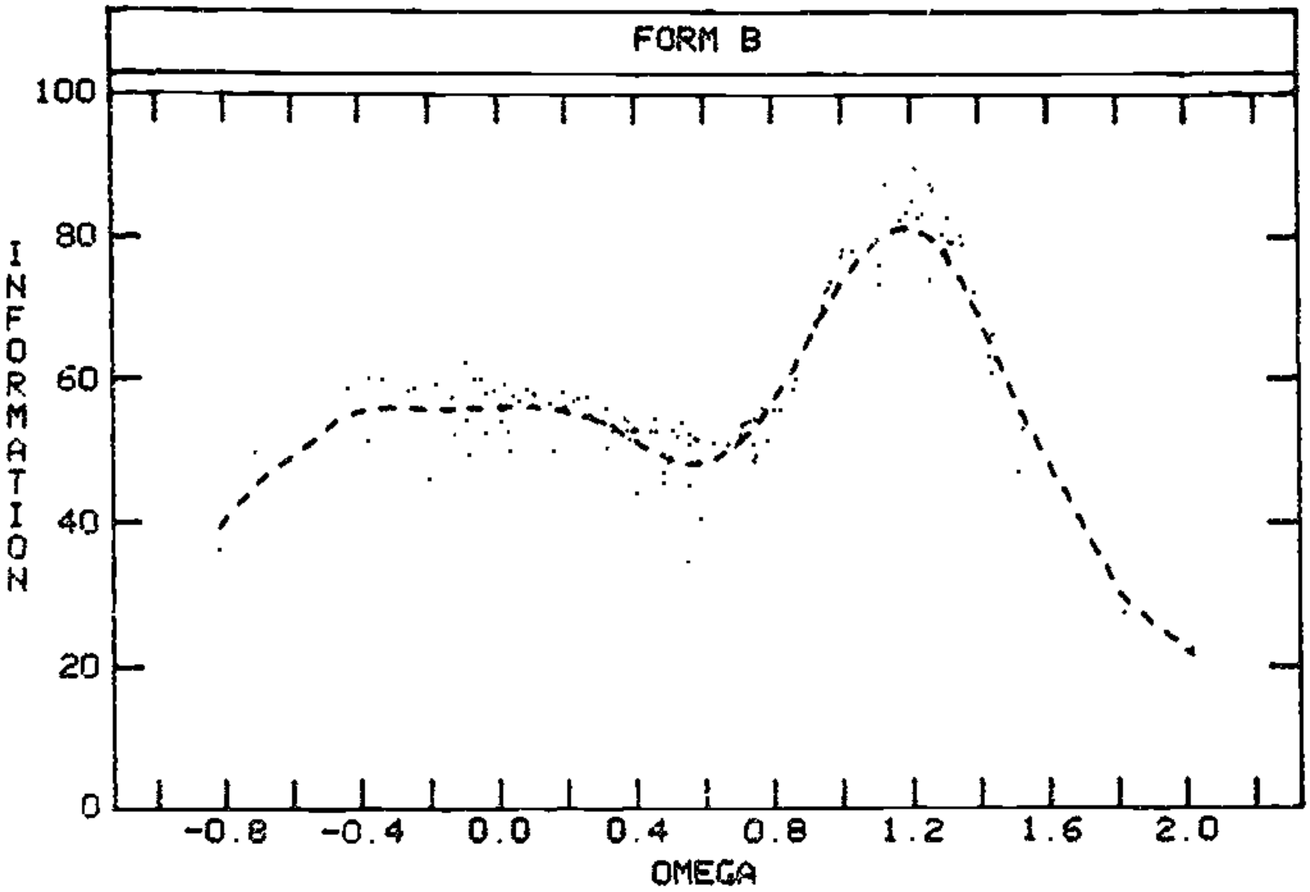


Figure 3.6b Observed information (dots), smoothed information, (Form B).

3.7 Parallelism, Reliability, and Validity at the 25th Item

This section presents data relevant to three important characteristics of the BRTT; the reliability of the scores, the validity of the construct measured, and the parallelism between each score formulated for Forms A and B of the test.

Table 3.7a presents reliabilities for seven scores obtained from the BRTT. The reliabilities are measured by the Pearson product moment correlation coefficient between the scores obtained on Form A and those obtained on Form B for each student. An adjusted reliability for the PSAT verbal score was computed by obtaining the test's reliability from the statistical analysis report (Form 3APT1) published by Educational Testing Service as a standard postadministration procedure. The reliability was adjusted for the obtained sample by Gulliksen's formula (1950, p. 114):

$$r_{xx} = 1 - \frac{S_x^2}{\sigma_x^2}(1 - \rho_{xx})$$

where:

ρ_{xx} is the reliability of the test for the population

σ_x^2 is the variance of the test for the population

S_x^2 is the variance for the sample.

The published reliability was $r_{xx} = .89$ with $\sigma_x^2 = 11.811$. Given sample variance $S_x^2 = 10.62$, the adjustment yielded a reliability $r_{xx} = .9111$.

To facilitate comparison of the 65 item PSAT with the 25 item BRTT, the Spearman-Brown formula

$$r' = \frac{K\, r_{xx}}{1 + (K-1)\, r_{xx}}$$

was applied with K=65/25. The expected reliability of the PSAT reduced to 25 items was 0.65. The reliability is directly comparable to the correlations shown for the BRTT scores.

Inspection of the column headed r_{ab} in Table 3.7a reveals that all of the BRTT scores were more reliable than the PSAT score at the 25th item. The highest reliability was found for Θ, Ω and $\underline{p}$. The scores which were computed from the mean difficulty of all items administered (b_{total}) and of all items answered correctly (b_{high}) or the final item administered (b_{final}) displayed the lowest reliability. In addition to the Pearson product moment correlation coefficient, Table 3.7a gives Spearman's rho which is a measure of reliability insensitive to a monotone transformation of the score. All findings based on these obtained values of the measure are compatible with those based on the Pearson product moment

Figure 3.7a is a scattergram of the Θ scores obtained for Form A vs. Form B. Figure 3.7b is a scattergram of Ω and Figure 3.7c is a scattergram of $\underline{p}$ (the expected proportion correct over the entire item pool). These figures include the five anomalous cases that exhibited peculiar response patterns. The three most separated points in the "northwest" corner of the scatterplots correspond to three of the anomalous cases. For these cases, Form B scores were considerably less than Form A scores. This discrepancy probably occurred because Form B was the second test and these individuals averaged less than eight seconds per item, suggesting a loss of attention or interest.

Table 3.7b summarizes the correlations between the scores from Forms A and B. These correlations and the Figures 3.7d, 3.7e, 3.7f and 3.7g indicate the ex.ent to which the scores preserve the ranking among individuals as ordered by θ. As can be seen from the graphs, Ω and p preserve order exactly. The scores that are based directly on the difficulties are b_{total}, $b_{correct}$, b_{high} and b_{final}. The relation between these and and the maximum likelihood estimate θ is depicted in the scatterplots of Figures 3.7h through 3.7o. As the scatterplots indicate, these scores are not simple monotone transformations of θ as are p and Ω, but are distinctly different from θ. The construct measured by these scores, and its relation to the theoretical ability θ, is an area for further research which should be conducted before a determination is reached concerning which scores should be used in practice.

Table 3.7b also presents correlations between the BRTT scores and the PSAT verbal score; Table 3.7c shows the frequency distribution of PSAT scores. Figures 3.7p through 3.7cc display the associated scatterplots. The correlations are adjusted for tests of equal length. These cor'elations may be interpreted as a form of concurrent validity; they indicate the extent to which the BRTT and PSAT measure a common construct. As can be seen from the table, the correlations betwee' the two tests were reasonably high, and $b_{correct}$ and b_{high} had among the highest correlations with the PSAT. Since a high score on the PSAT requires that the student answer a large number of items correctly, including some of high difficulty, it is possible that this relationship results from a psychological variable related to accuracy on difficult items. This explanation is highly speculative, but it offers an intriguing possibility for future research. Although the BRTT-PSAT correlations are high, they are not perfect. This suggests that the two tests do not measure exactly the same construct. The difference may occur because the PSAT score includes items which measure reading comprehension skills while the BRTT does not include such items. Also, the BRTT is computer-presented while the PSAT is a pencil-and-paper instrument. Finally, the variation may be partly attributed to the usual differences which characterize items selected for any set of parallel tests.

An important question is that of the parallelism of the two forms of the BRTT. Parallelism in this context involves whether a given score from Form A was significantly different from the score as computed using Form B. If the two forms yielded different mean values with respect to a given score, the tests would have to be equated before individual score comparisons could be made.

Table 3.7d presents paired _t_ tests and one-sample van der Waerden tests between Forms A and B for the various scores. A significant test statistic indicates that the differences between scores were not due to chance, and the two forms cannot be considered parallel with respect to the score. All eleven anomalous cases were omitted in these analyses. The one-sample van der Waerden test was performed to insure that any significant t statistic was due to a real difference between Forms A and B and not an artifact due to non-normality in the data.

The data presented in this section indicates that _p_, or a score closely related to it, would be the best choice for the BRTT. In terms of reliability, _p_ ranks, with θ and [illegible], among the most reliable of the scores studied. Unlike θ and [illegible], _p_ does not suffer from the infinity problem discussed in Section 3.9. A student who answers all items incorrectly on the BRTT would obtain a $\underline{p}$ score of 0; one who answers all items correctly would obtain a $\underline{p}$ score of 1. In the case of θ and Ω, such individuals would obtain inderterminant scores. Furthermore, for this data, $\underline{p}$ is a parallel score; whereas θ and Ω are not. This would indicate that the need for test equating is reduced when $\underline{p}$ is used providing the item pools for the alternative forms are comparable with respect to the a, b, and c parameters.

Table 3.7a

Correlation Between Forms A and B

for Score at 25th Item ($\underline{N}$=135)

SCORE	$\underline{r}_{ab}$	$\underline{rho}_{ab}$
Θ	.8719	.8585
Ω	.8730	.8585
b_{total}	.8247	.8163
$b_{correct}$	.8195	.8068
b_{high}	.7261	.7448
b_{final}	.6985	.6508
p	.8732	.8585

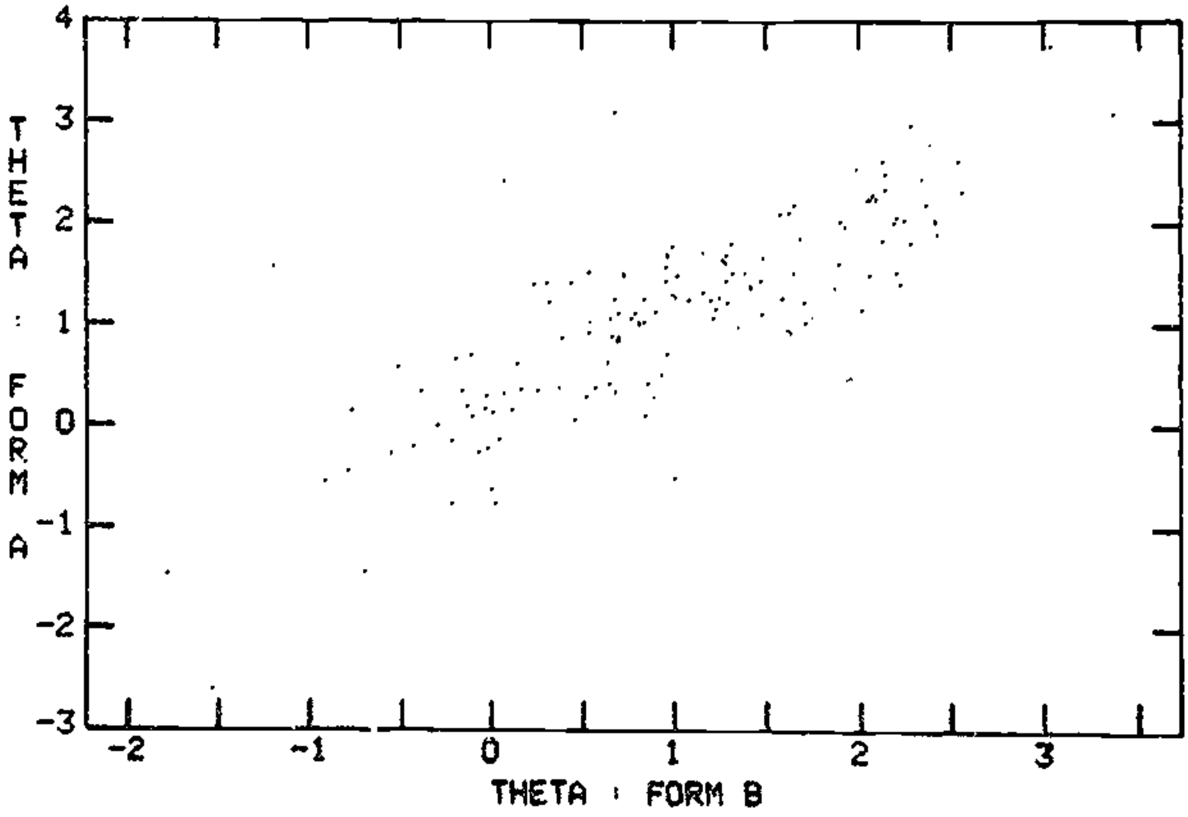


Figure 3.7a Scattergram of Theta (Form A) vs. Theta (Form B)

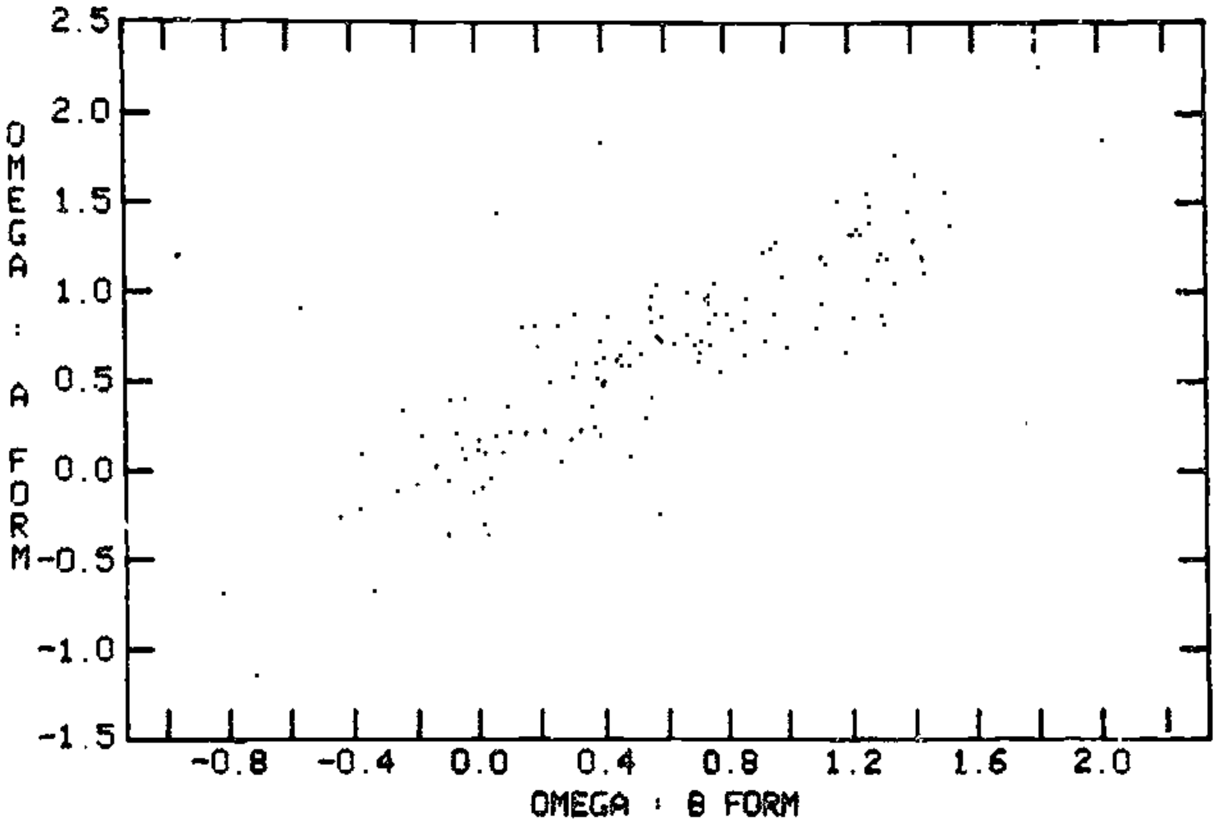


Figure 3.7b Scattergram of Omega (Form A) vs. Omega (Form B)

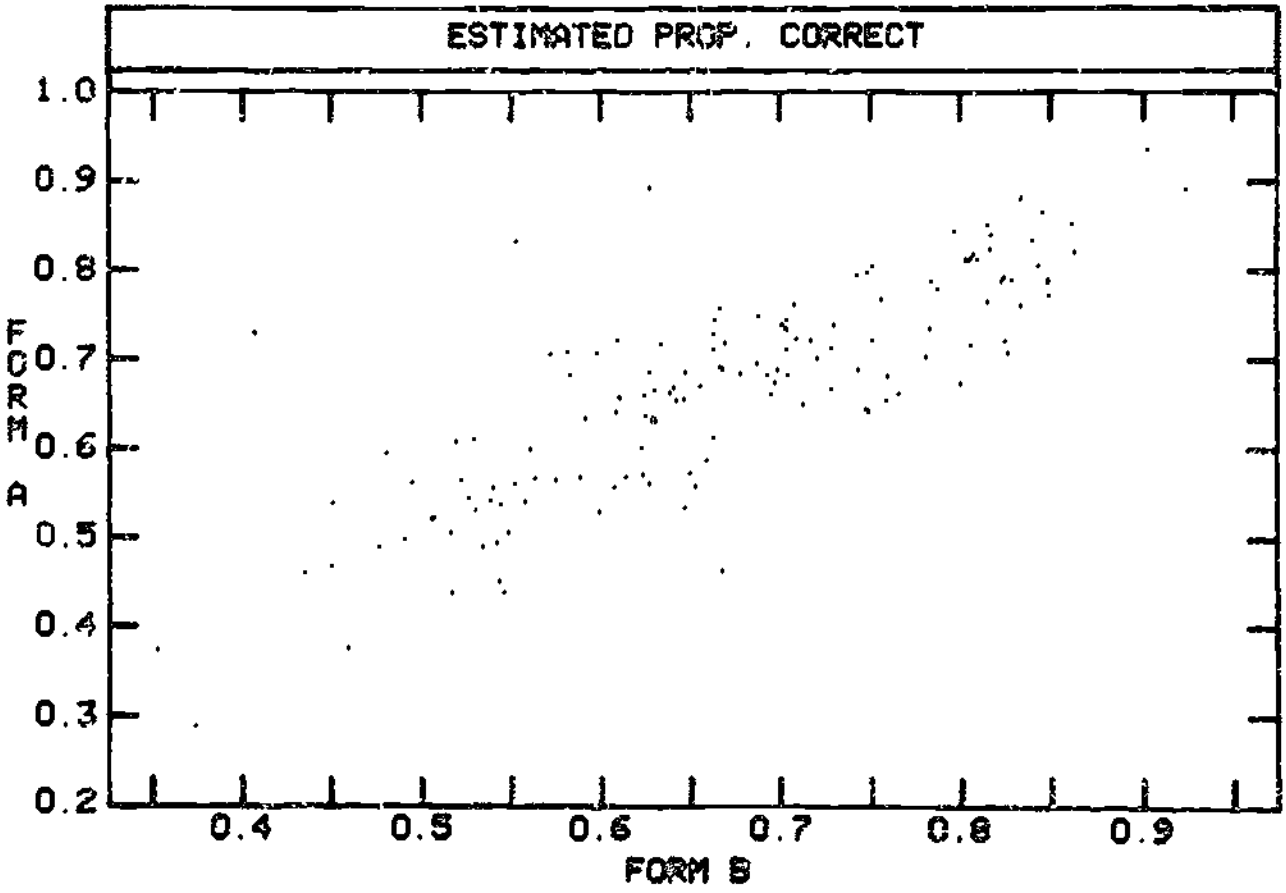


Figure 3.7 Scattergram of p_a vs. p_b

Table 3.7b

Correlations Between Score at 25th Item, θ, and PSAT Verbal

Score	$r_{a\theta}$ N=135	$r_{b\theta}$ N=135	$r_{a\ \mathrm{PSAT}}$[†] N=92	$r_{b\ \mathrm{PSAT}}$[†] N=92
θ			.8517 (.7547)	.8745 (.7749)
Ω	.9978	.9987	.8616 (.7684)	.8750 (.7803)
b_{total}	.9717	.9781	.8335 (.7225)	.8798 (.7626)
$b_{correct}$	.9695	.9778	.8323 (.7191)	.8848 (.7645)
b_{high}	.9503	.8828	.9105 (.7405)	.9042 (.7354)
b_{final}	.8449	.9144	.8303 (.6623)	.8655 (.6904)
p	.9954	.9989	.8575 (.7648)	.8719 (.7776)

[†] Adjusted for attenuation (see Lord and Novick (1967, p.70)).
The number appearing in parenthesis is the unadjusted Pearson correlation coefficient.

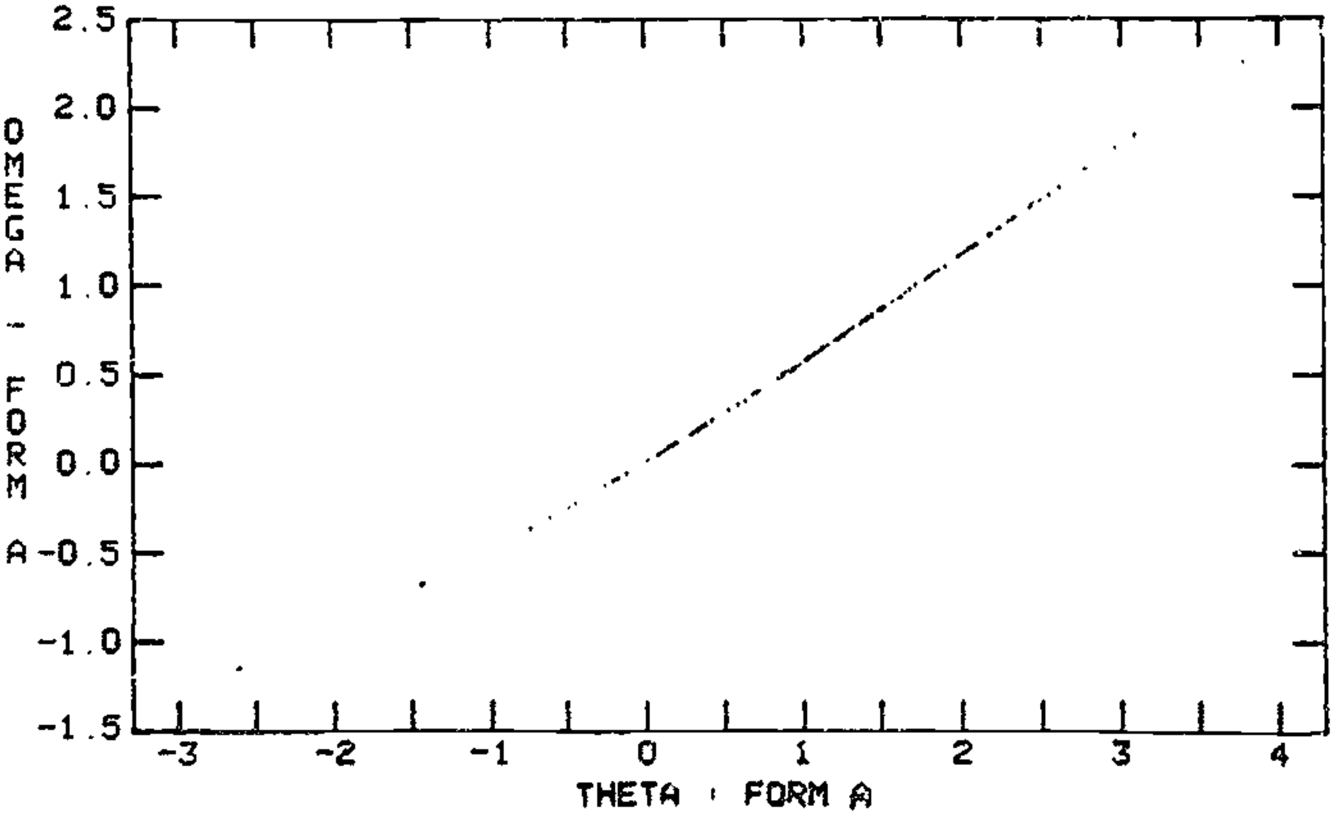


Figure 3.7d Omega vs. Theta (Form A)

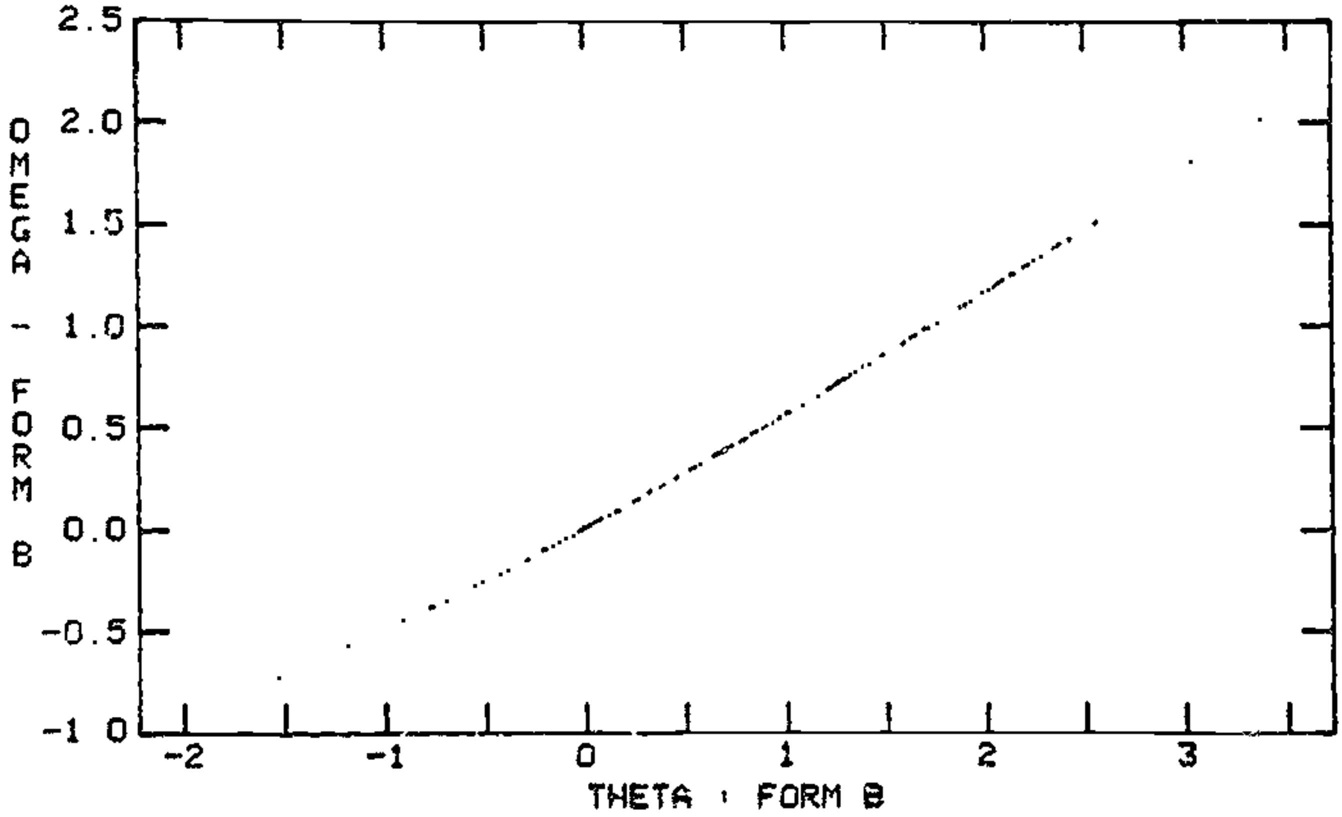


Figure 3.7e Omega vs. Theta (Form B)

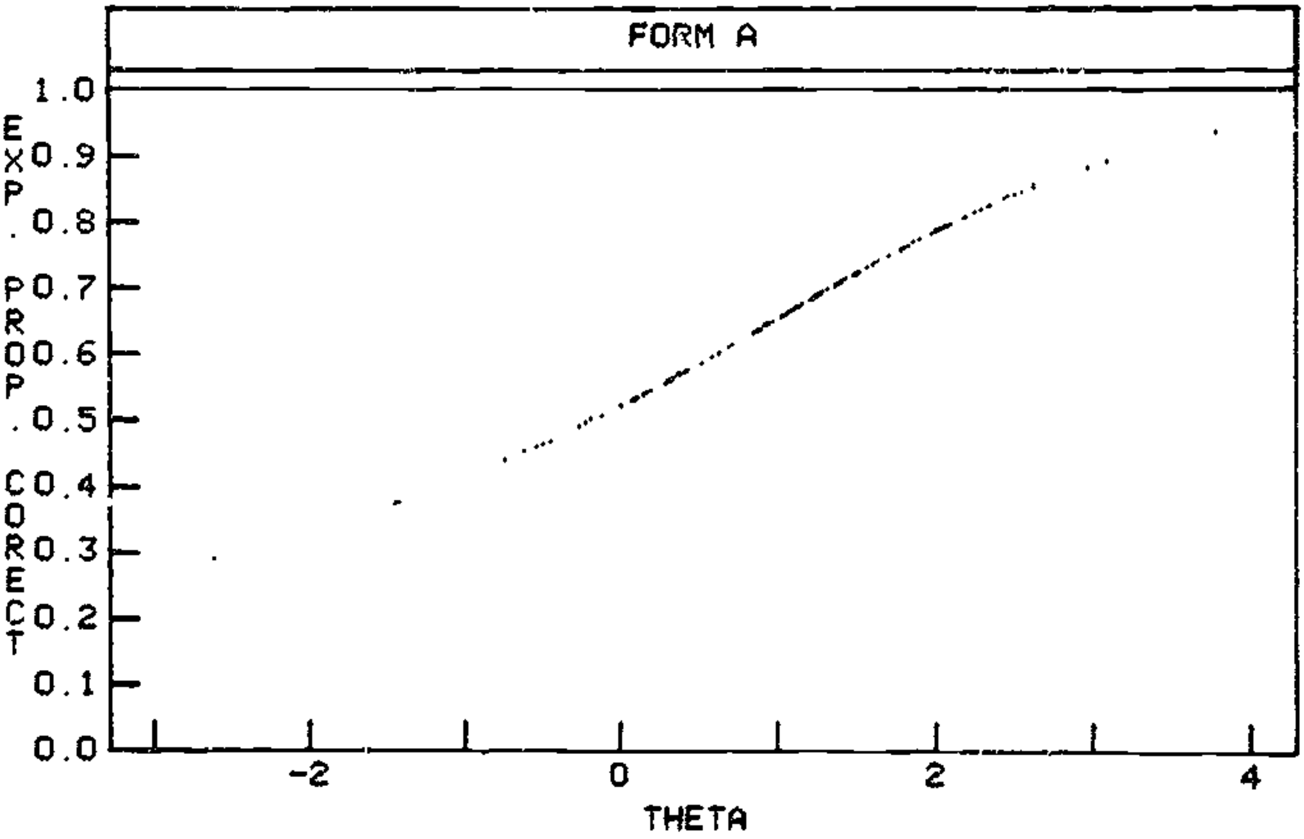


Figure 3.7f p vs. Theta (Form A)

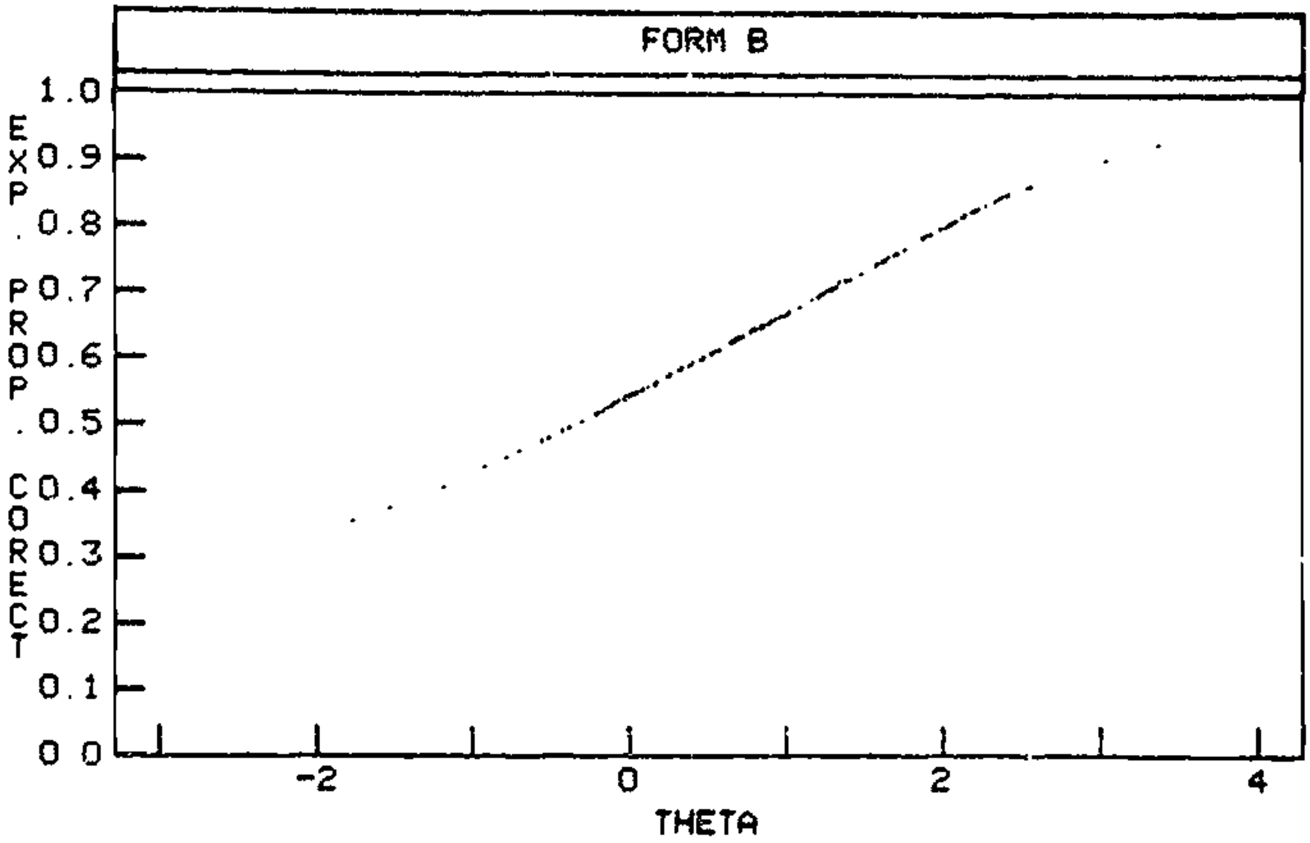


Figure 3.7g $\underline{p}$ vs. Theta (Form B)

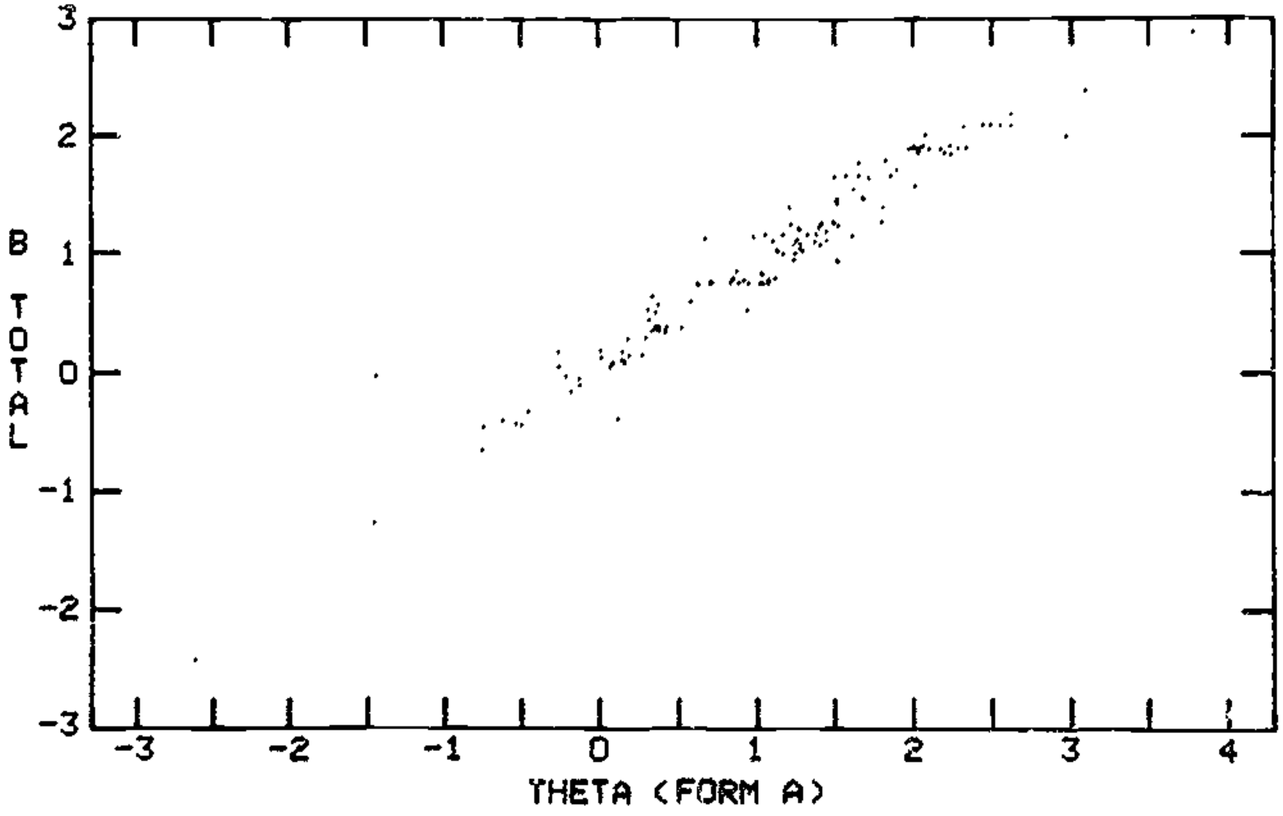
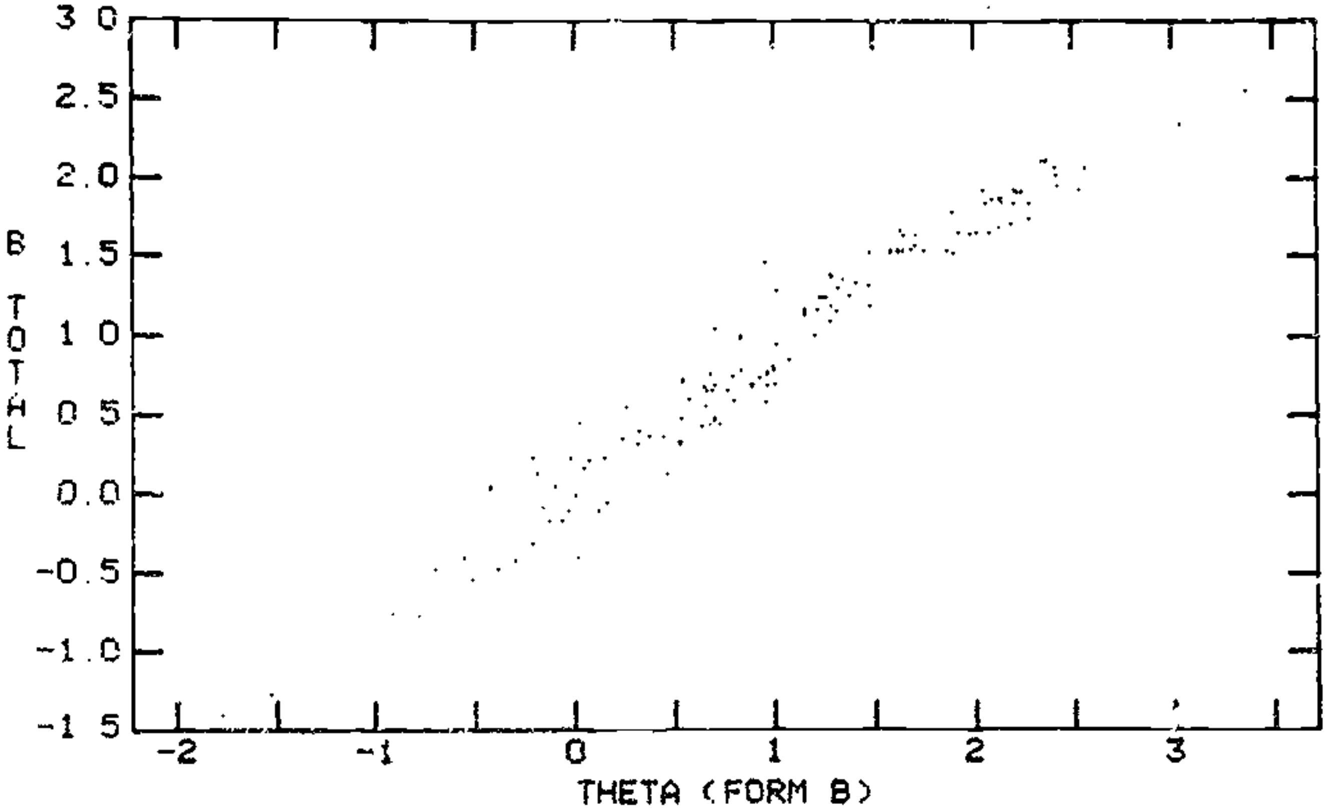


Figure 3.7h b_{total} vs. Theta (Form A)



F'gure 3.7i b_{total} vs. Theta (Form B)

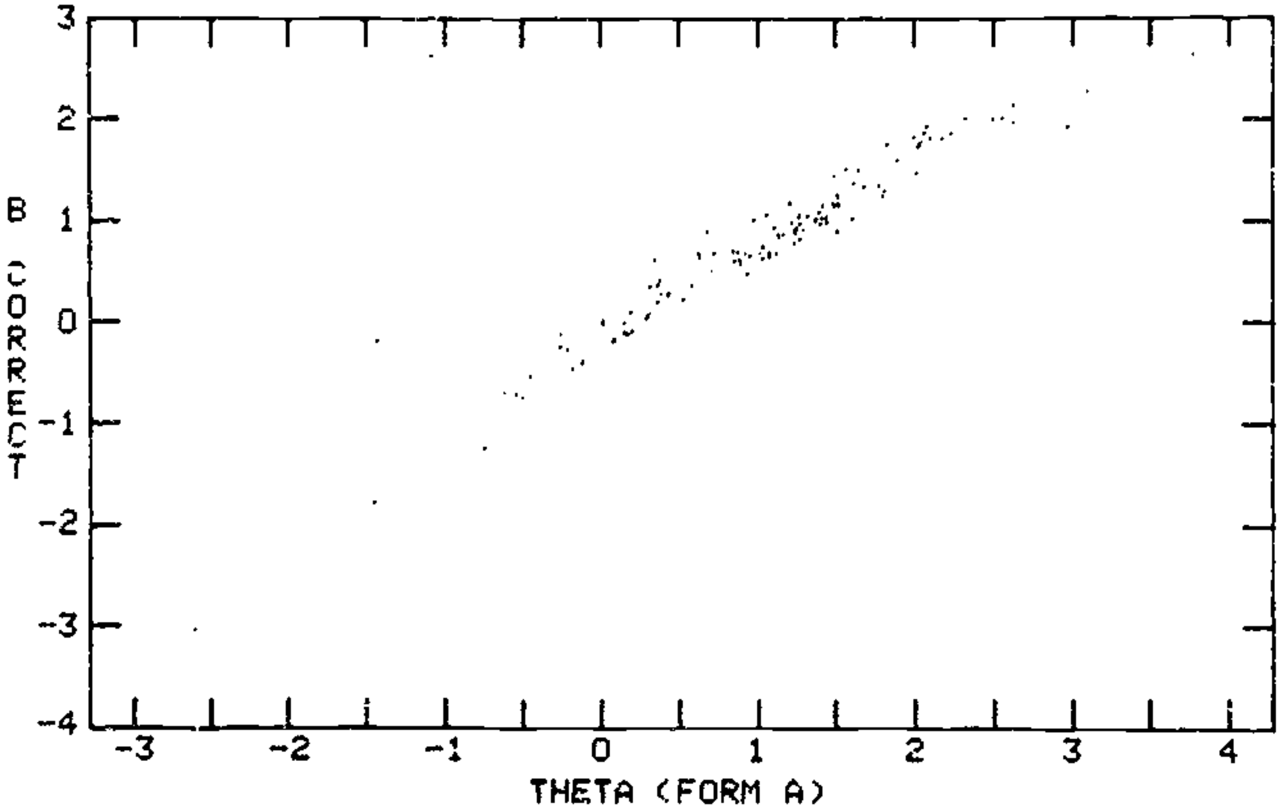


Figure 3.7j $b_{correct}$ vs. Theta (Form A)

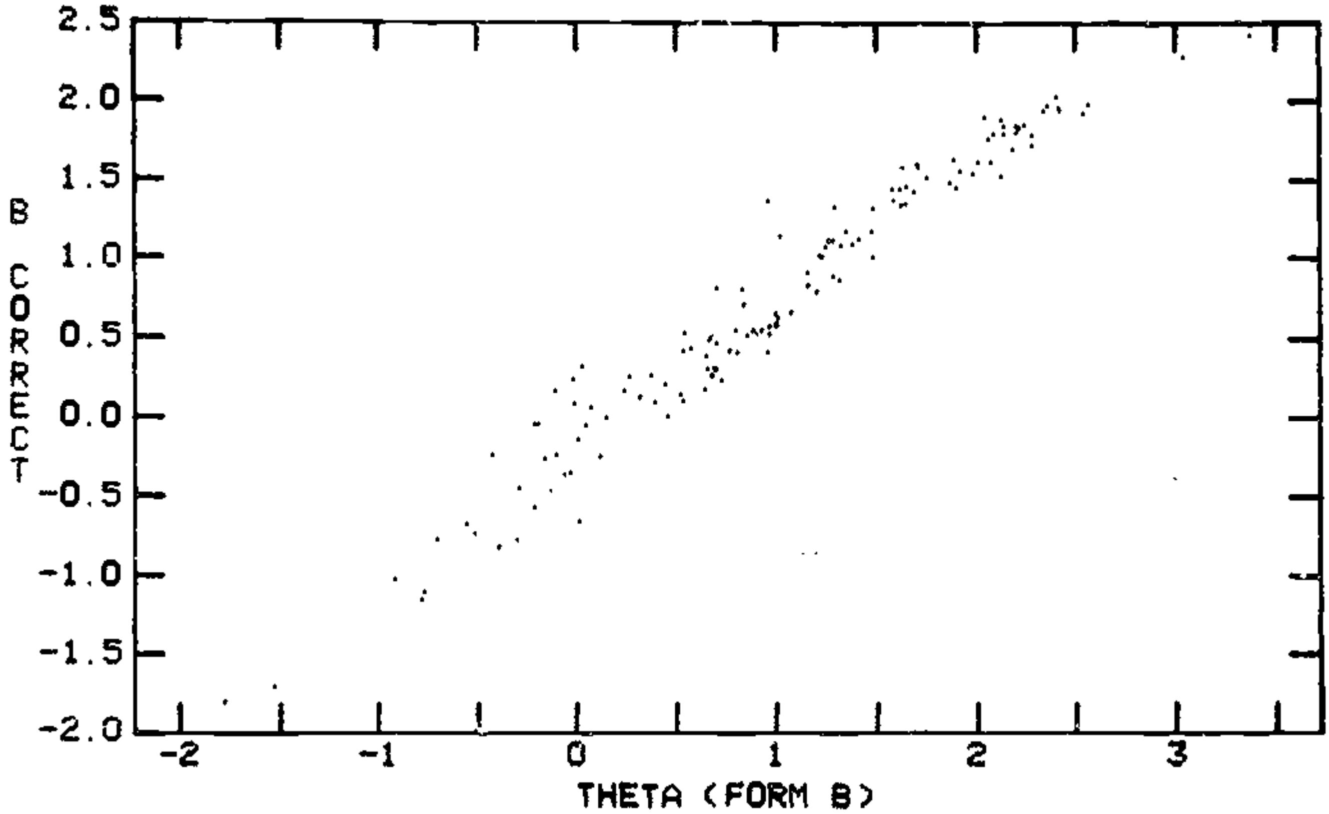


Figure 3.7k $b_{correct}$ vs. Theta (Form B)

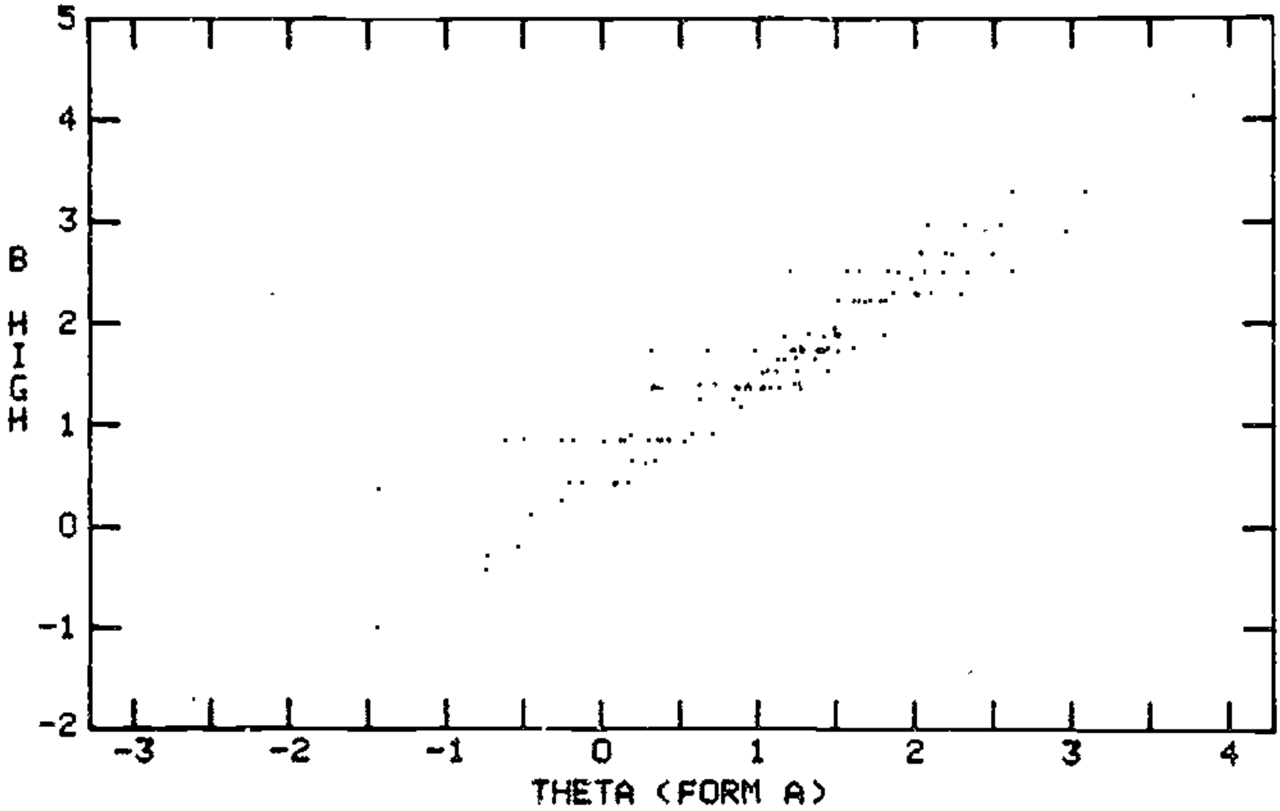


Figure 3.71 b_{high} vs. Theta (Form A)

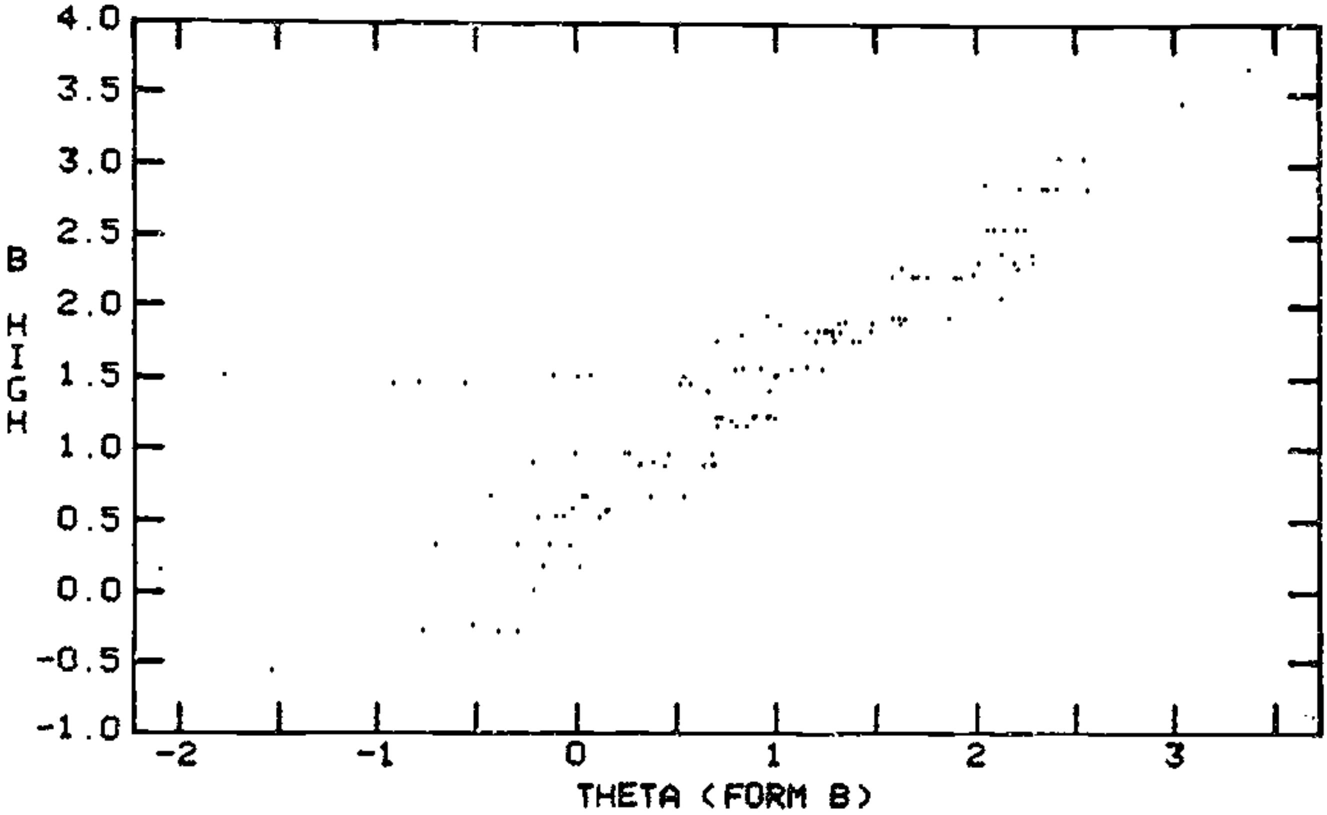


Figure 3.7m b_{high} vs. Theta (Form B)

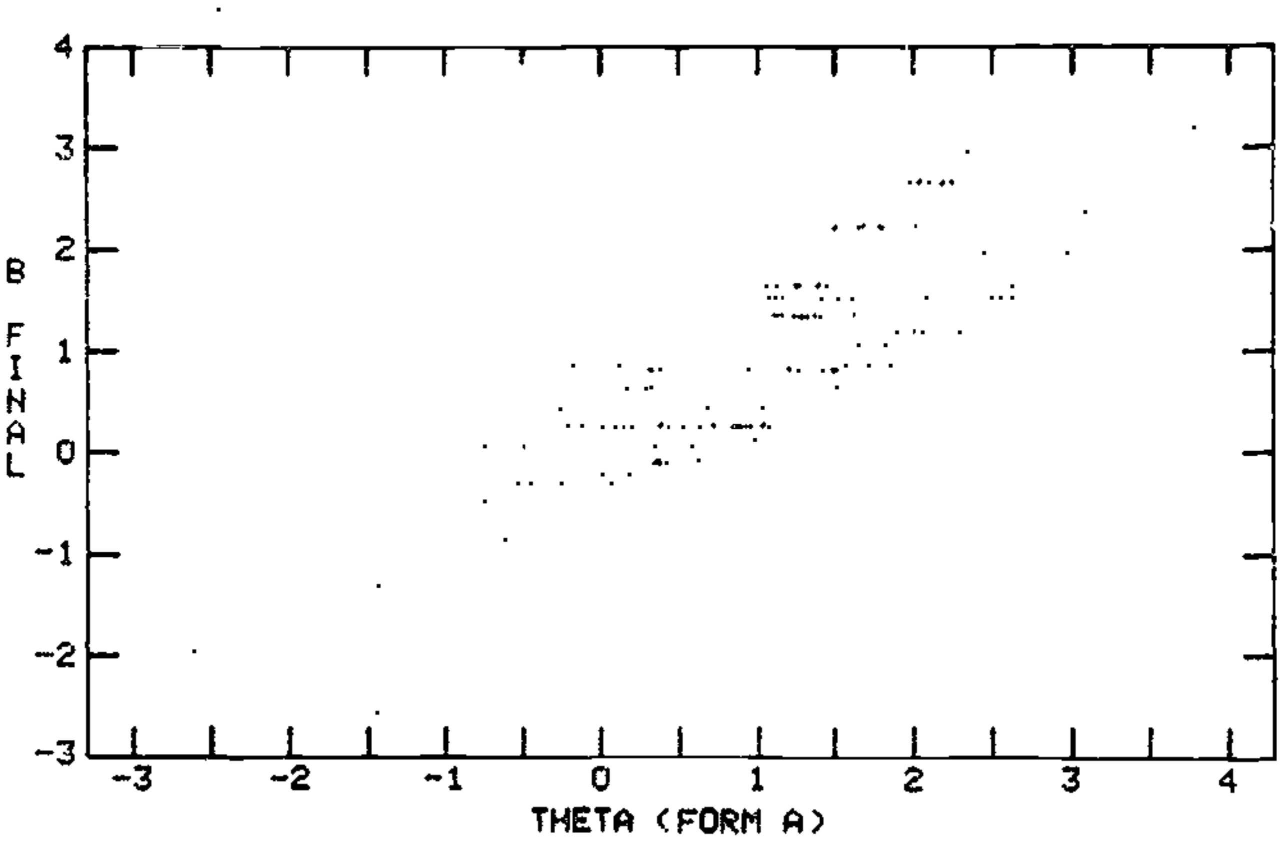


Figure 3.7n b_{final} vs. Theta (Form A)

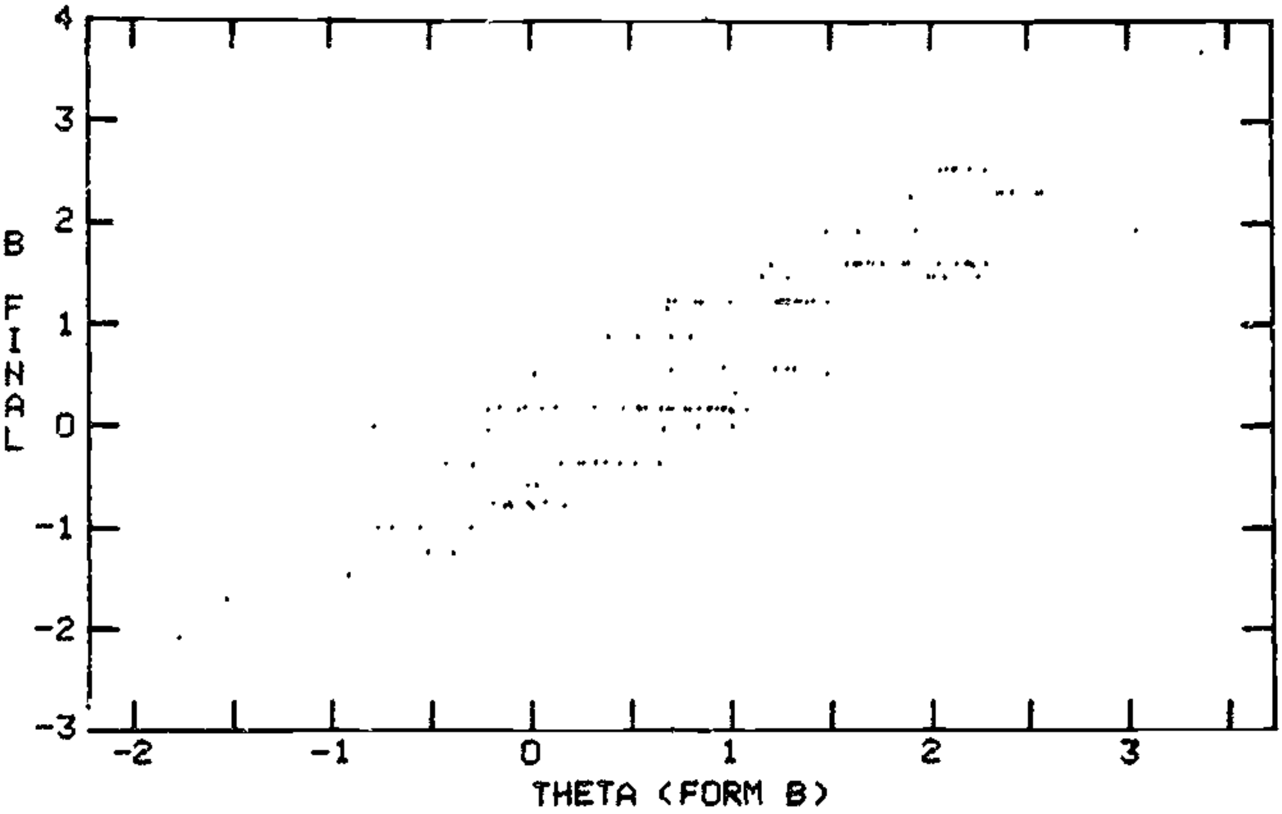


Figure 3.7o b_{final} vs. Theta (Form B)

Table 3.7c

Distribution of PSAT Scores

Score Interval	Frequency	%
71.5 - 76.0	1	1.1
66.5 - 71.5	9	9.8
61.5 - 66.5	8	8.7
56.5 - 61.5	4	4.3
51.5 - 56.5	14	15.2
46.5 - 51.5	14	15.2
41.5 - 46.5	10	10.9
36.5 - 41.5	17	18.5
31.5 - 36.5	3	3.3
26.5 - 31.5	7	7.6
21.5 - 26.5	3	3.3
16.5 - 21.5	0	0.0
12.0 - 16.5	2	212

N	92
MEAN	48.177
SD	13.343
Minimum Value	12.854
Maximum Value	75.284

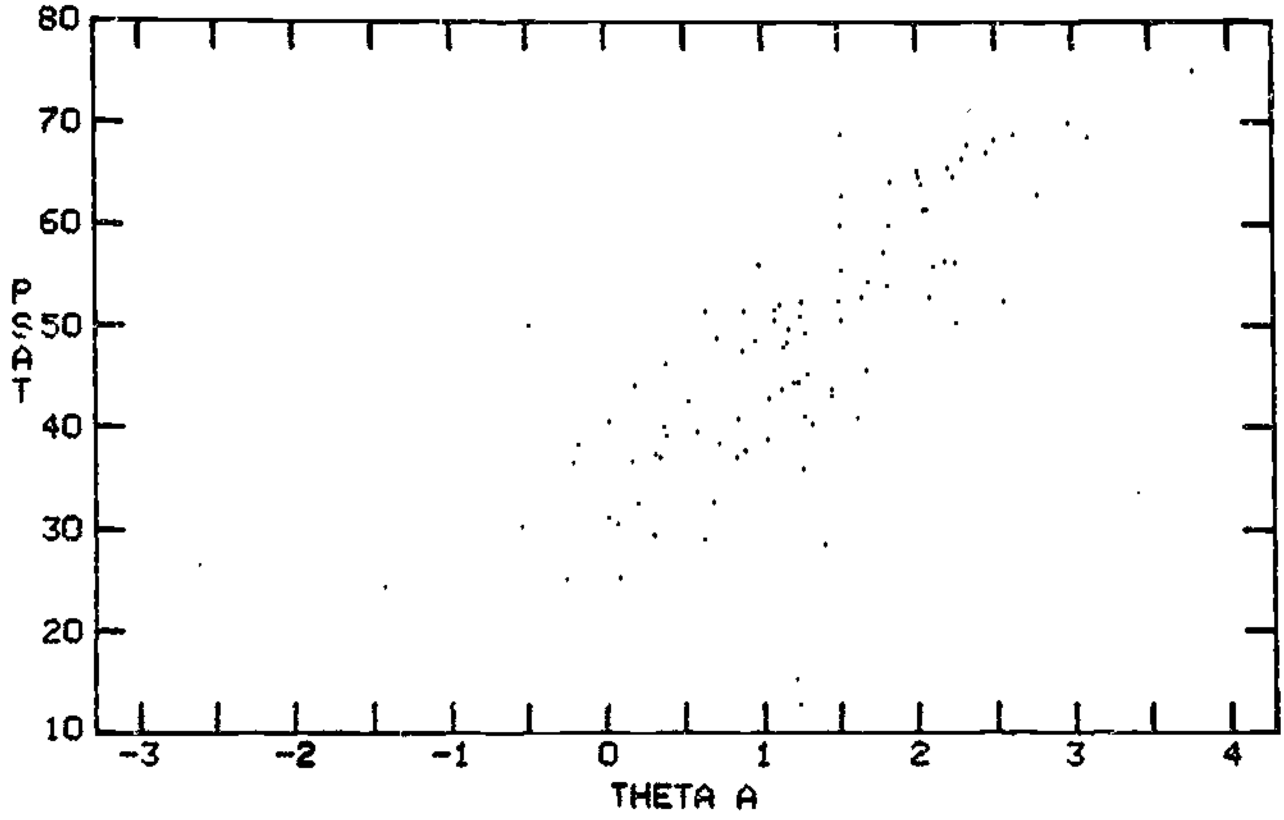


Figure 3.7p   PSAT vs. Theta (Form A)

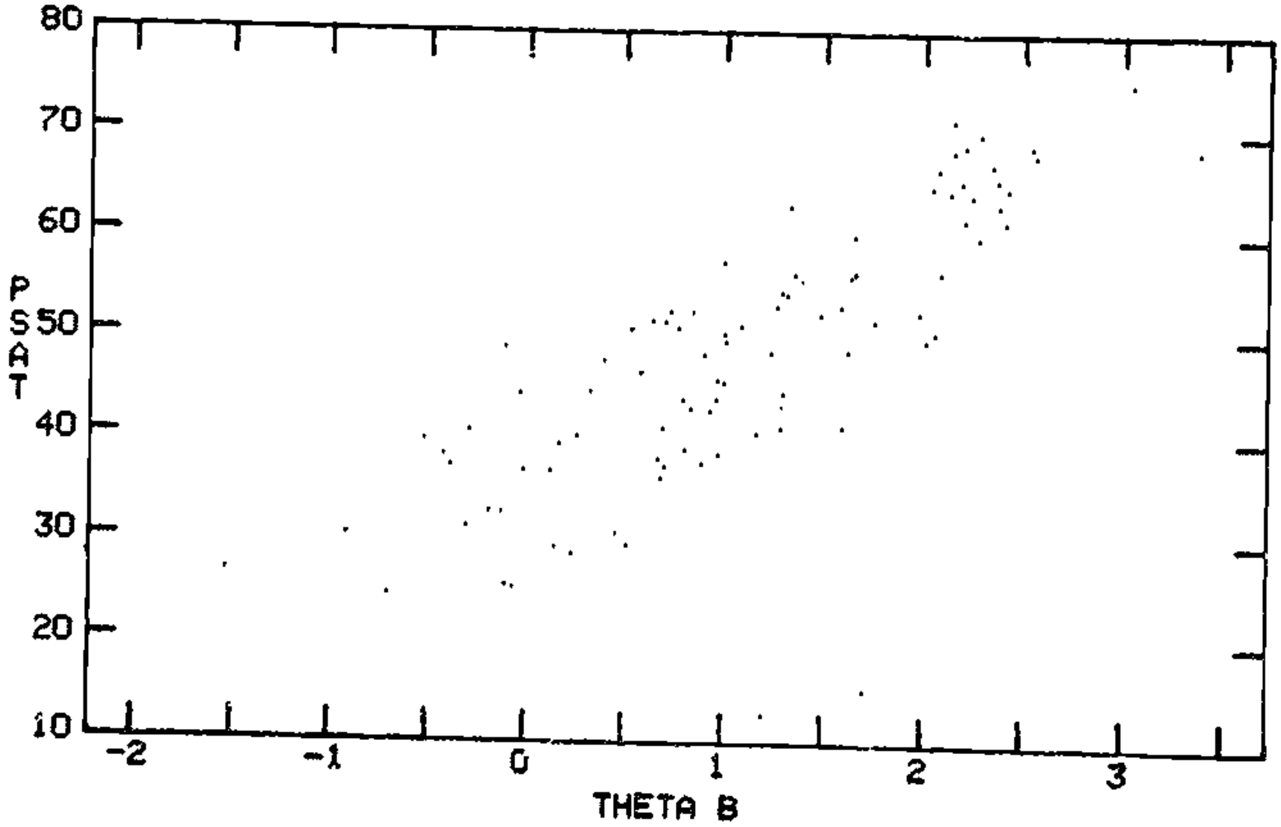


Figure 3.7q PSAT vs. Theta (Form B)

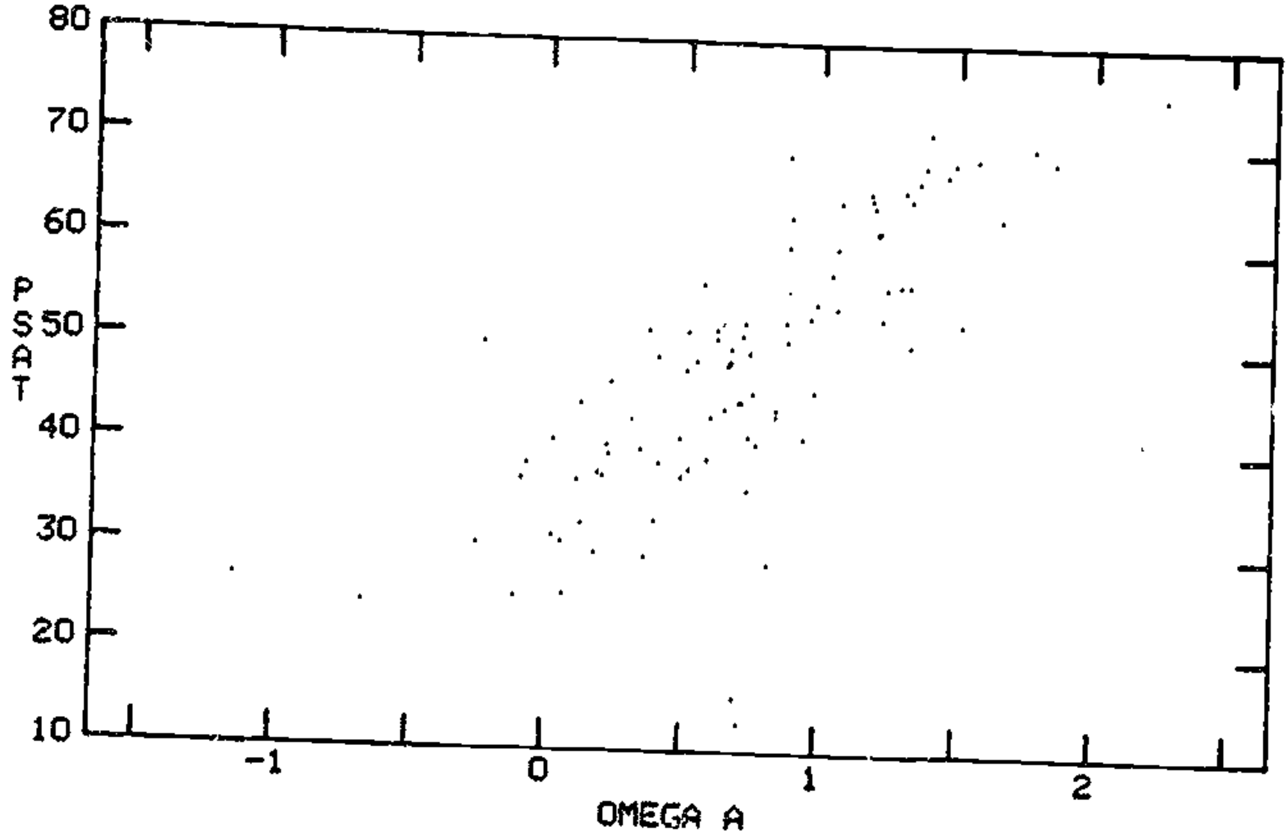


Figure 3.7r PSAT vs. Omega (Form A)

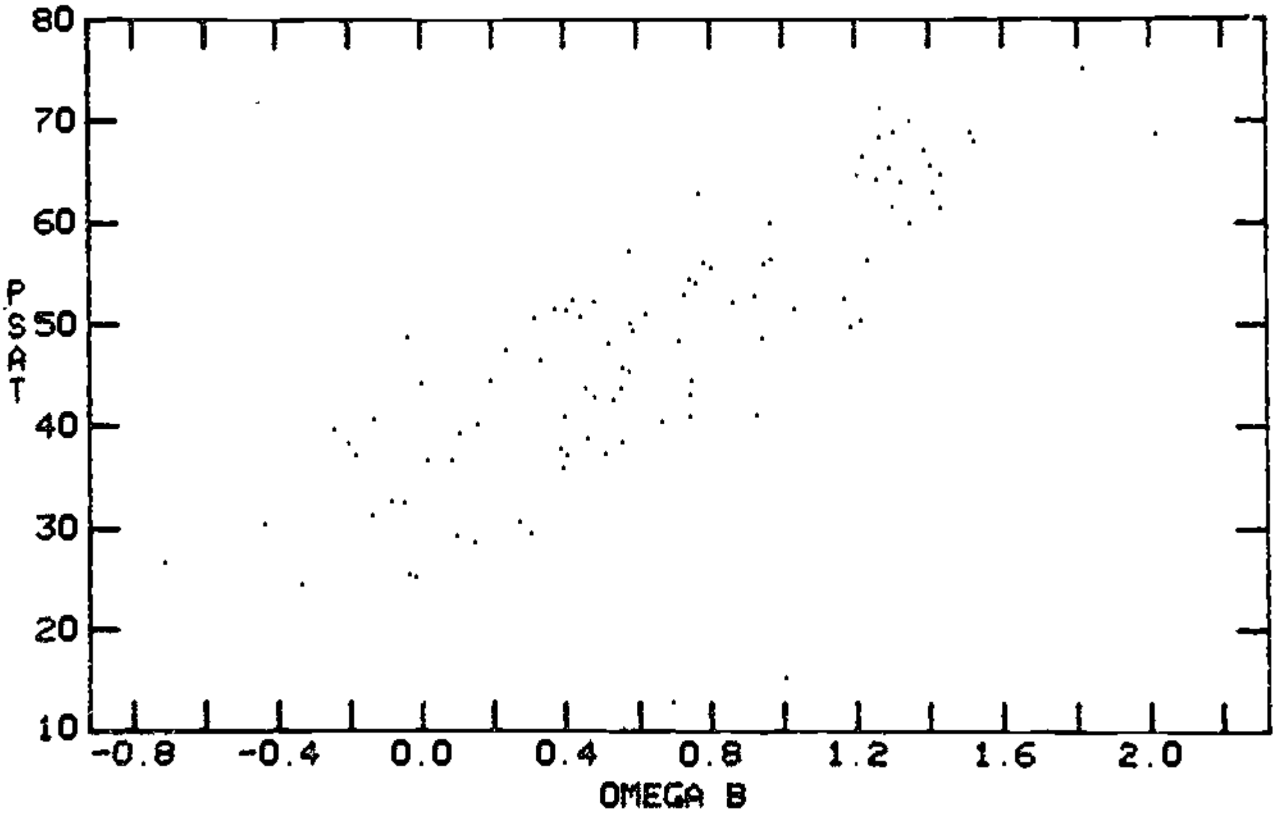


Figure 3.7s PSAT vs. Omega (Form B)

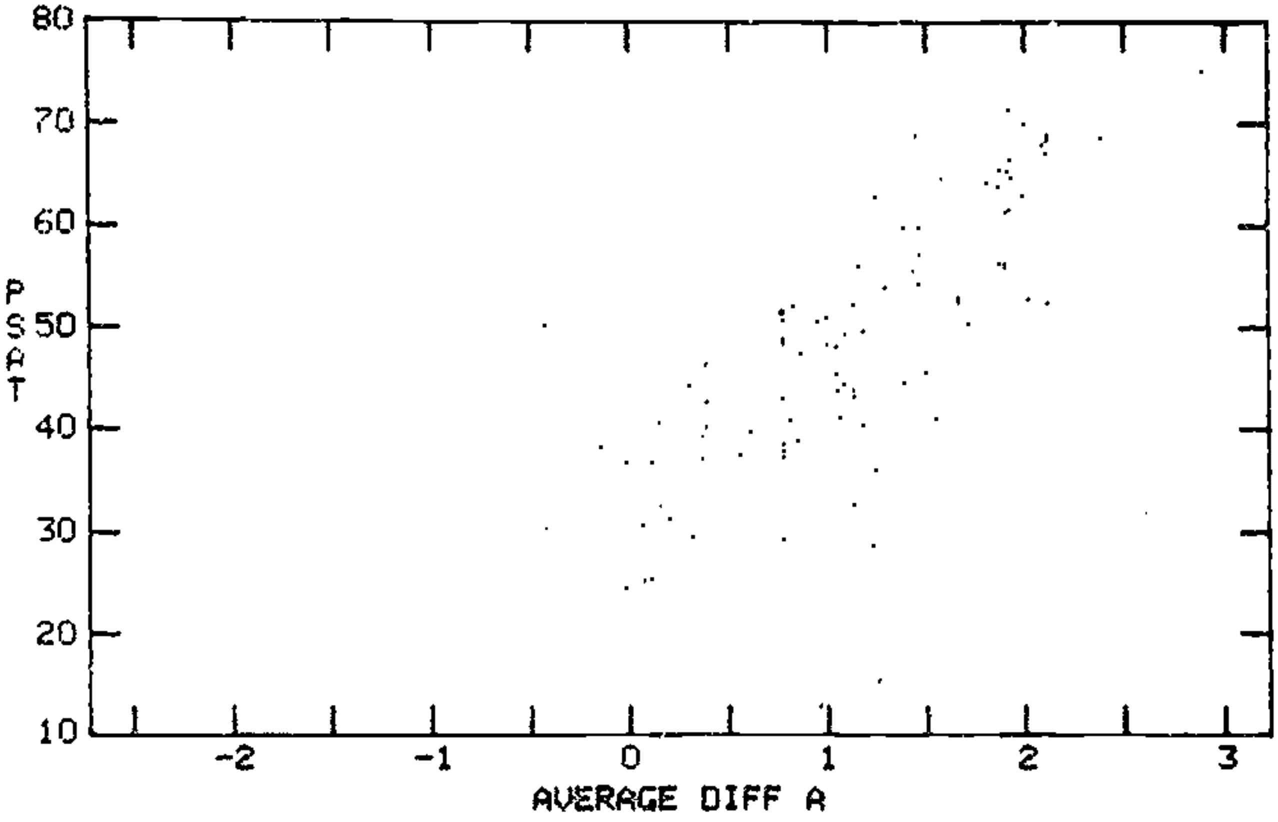


Figure 3.7t PSAT vs. b_{total} (Form A)

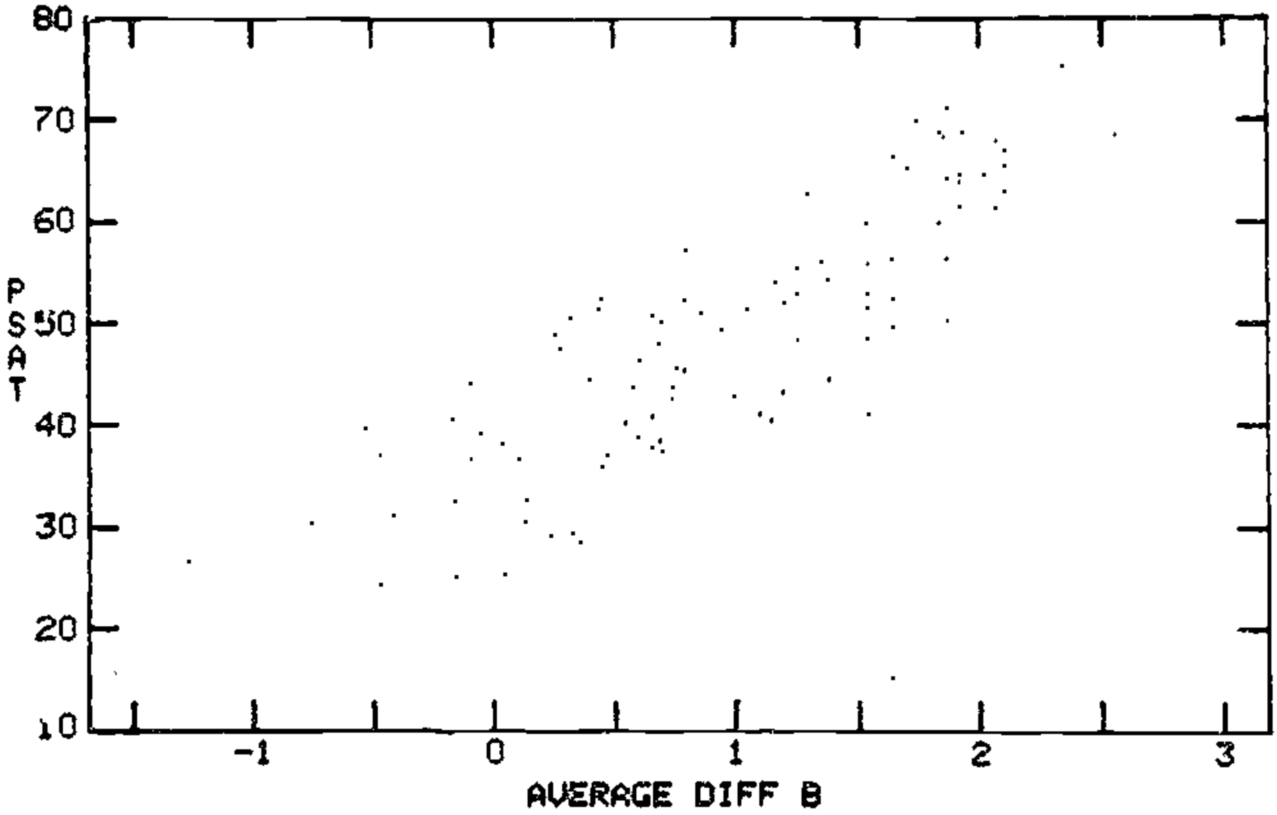
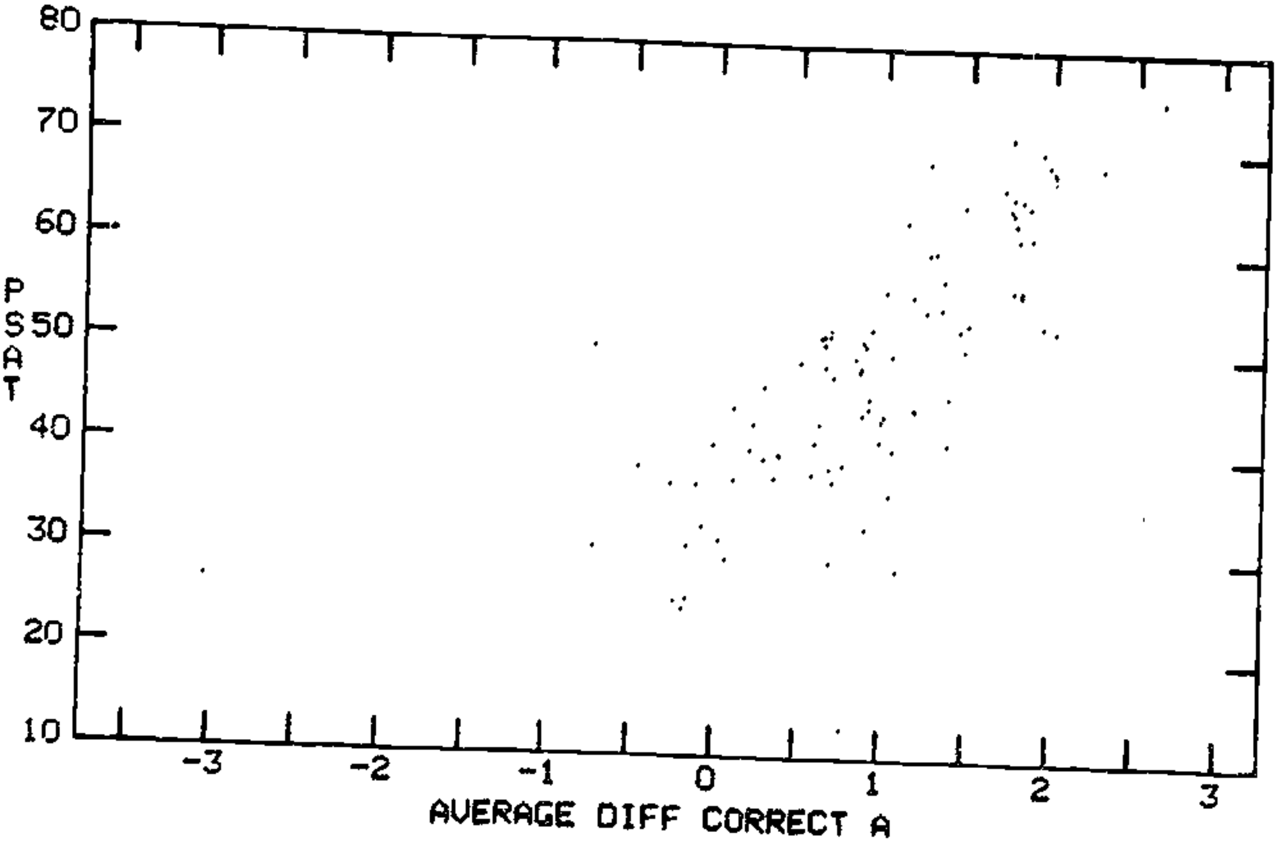


Figure 3.7c PSAT vs. b_{total} (Form B)



Figu.e 3.7v PSAT vs. $b_{correct}$ (Form A)

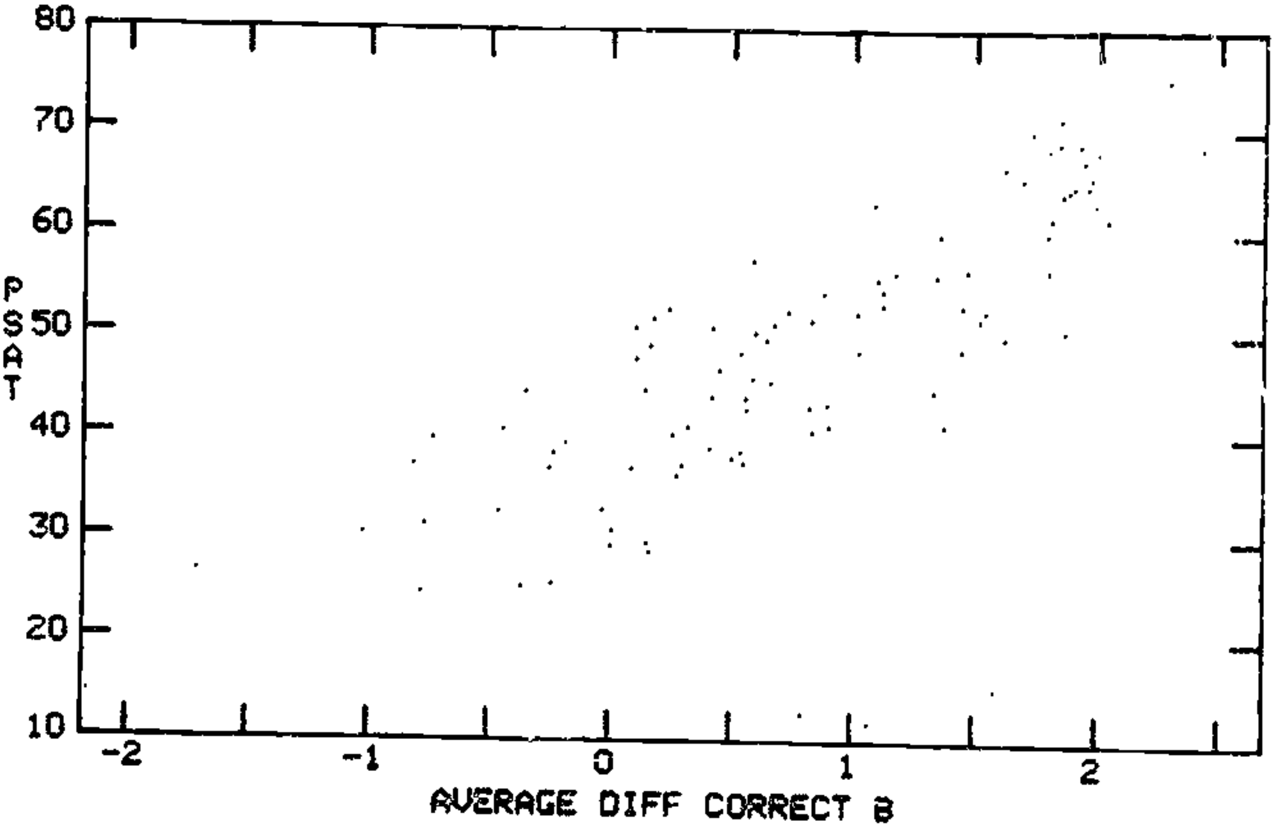


Figure 3.7w PSAT vs. $b_{correct}$ (Form B)

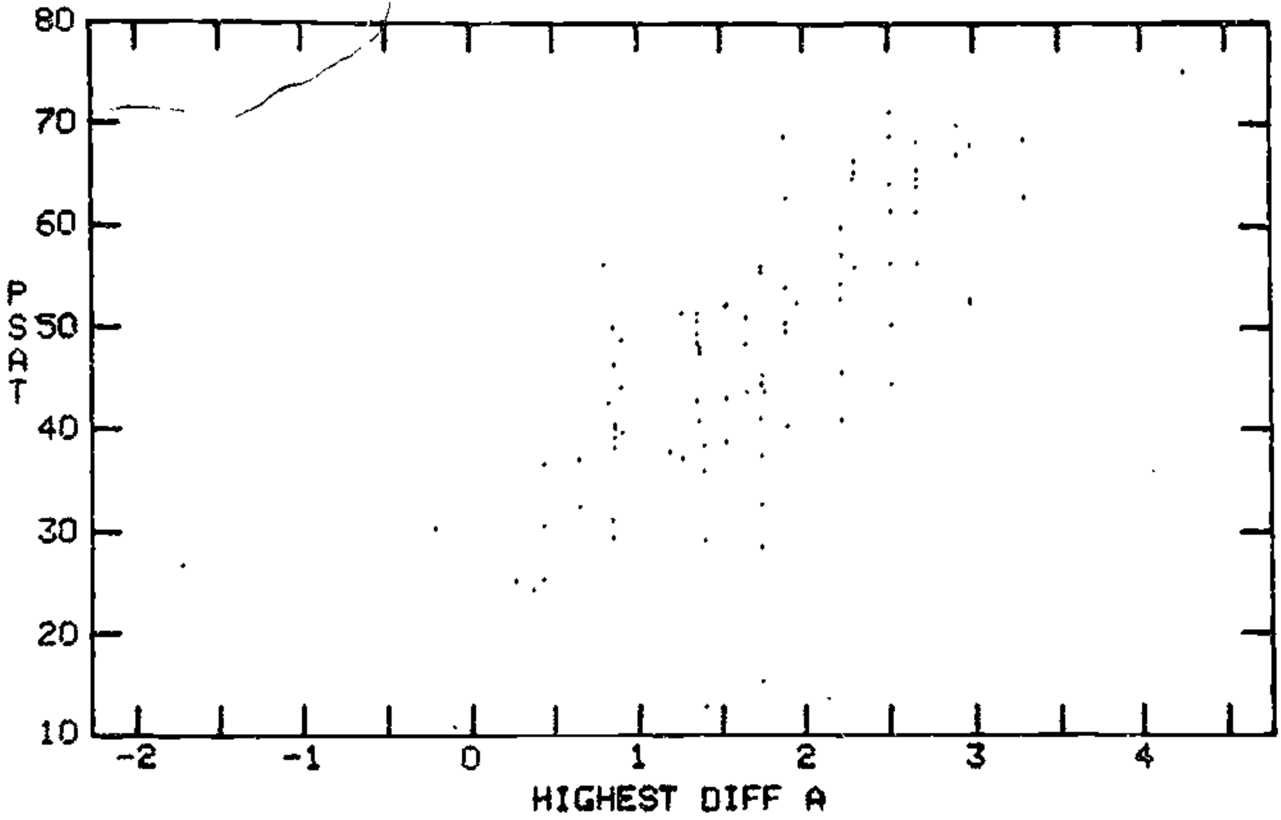


Figure 3.7x PSAT vs. b_{high} (Form A)

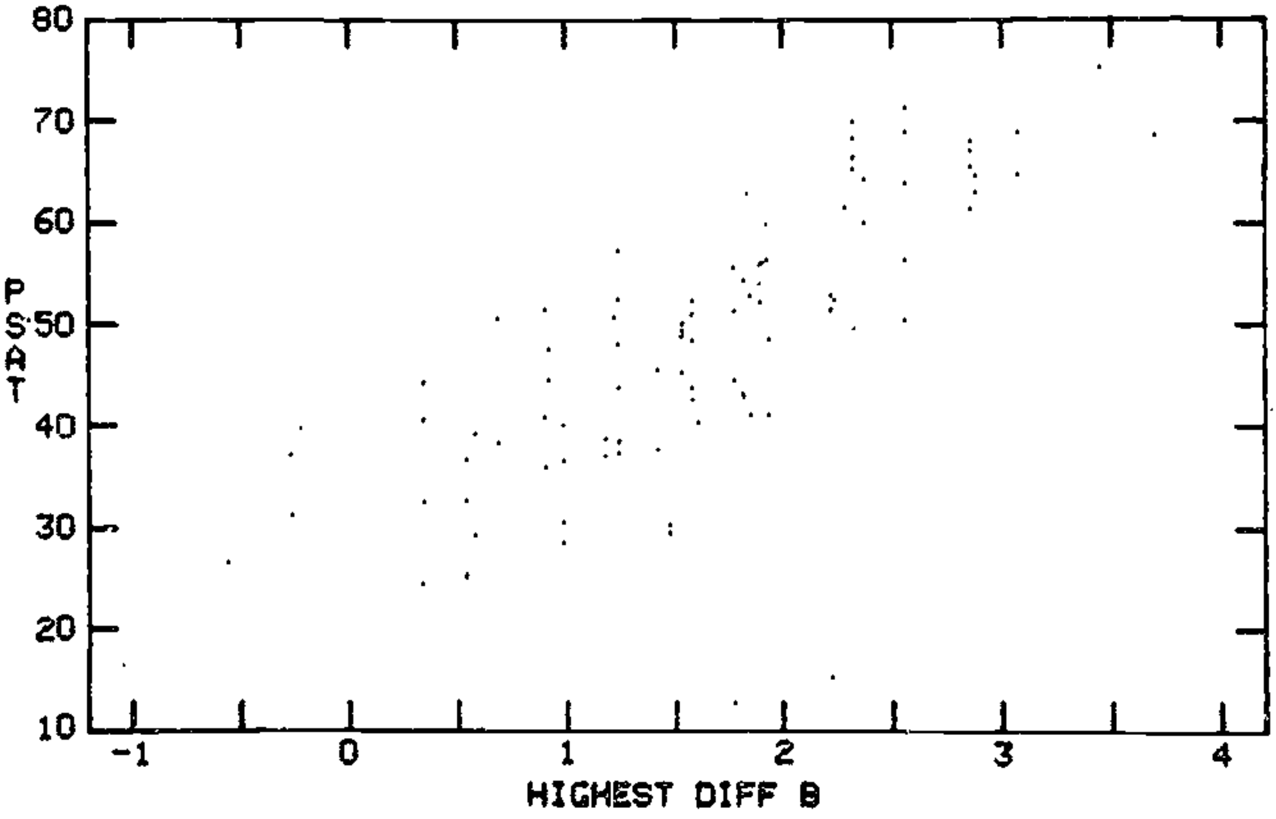


Figure 3.7y PSAT vs. b_{high} (Form B)

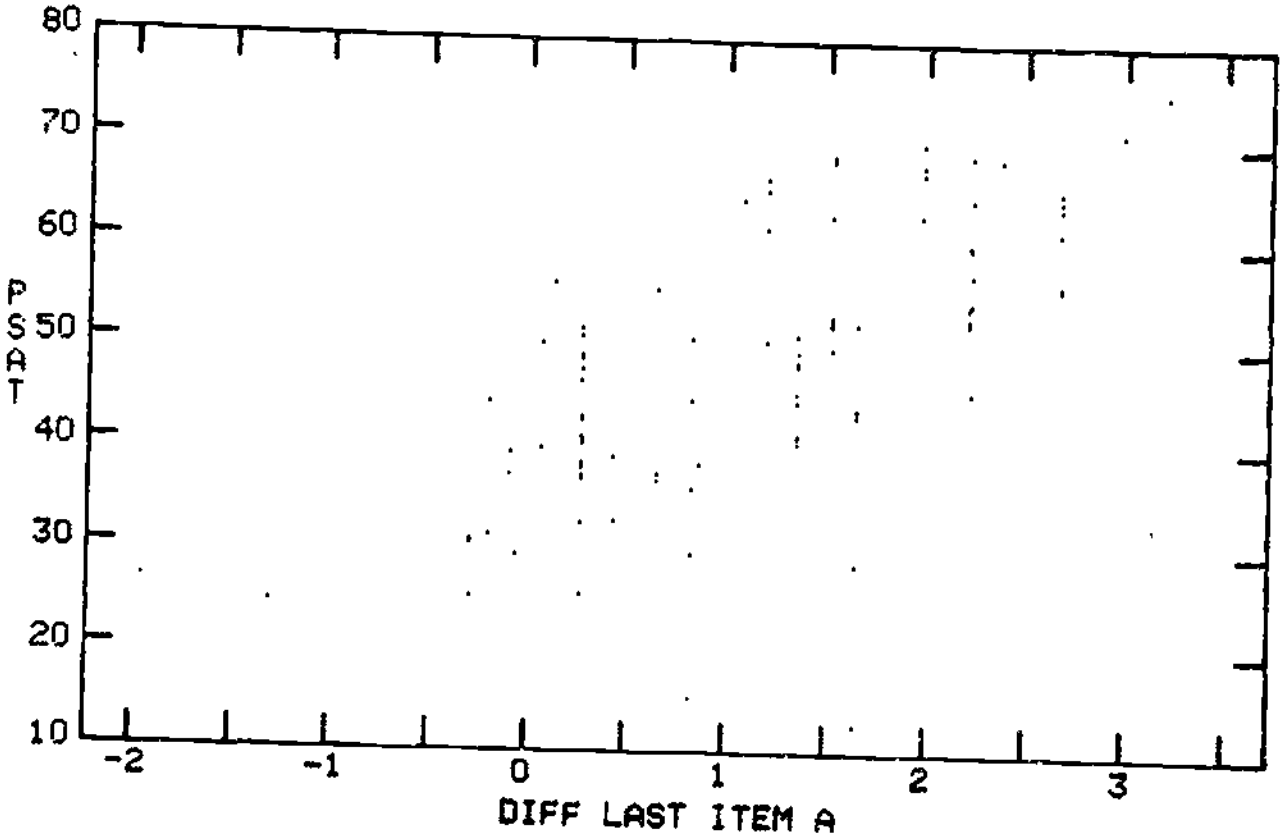


Figure 3.7z PSAT vs. b_{final} (Form A)

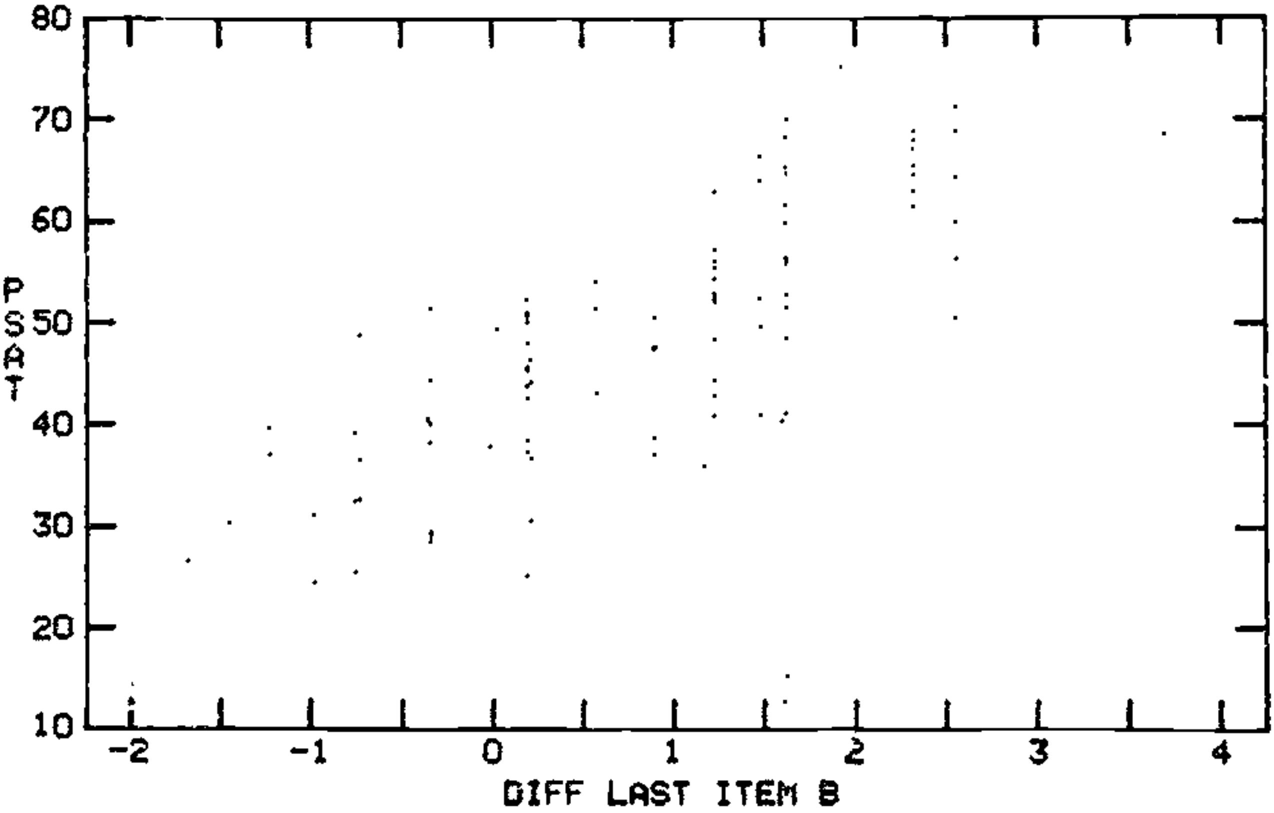


Figure 3.7aa PSAT vs. b_{final} (Form B)

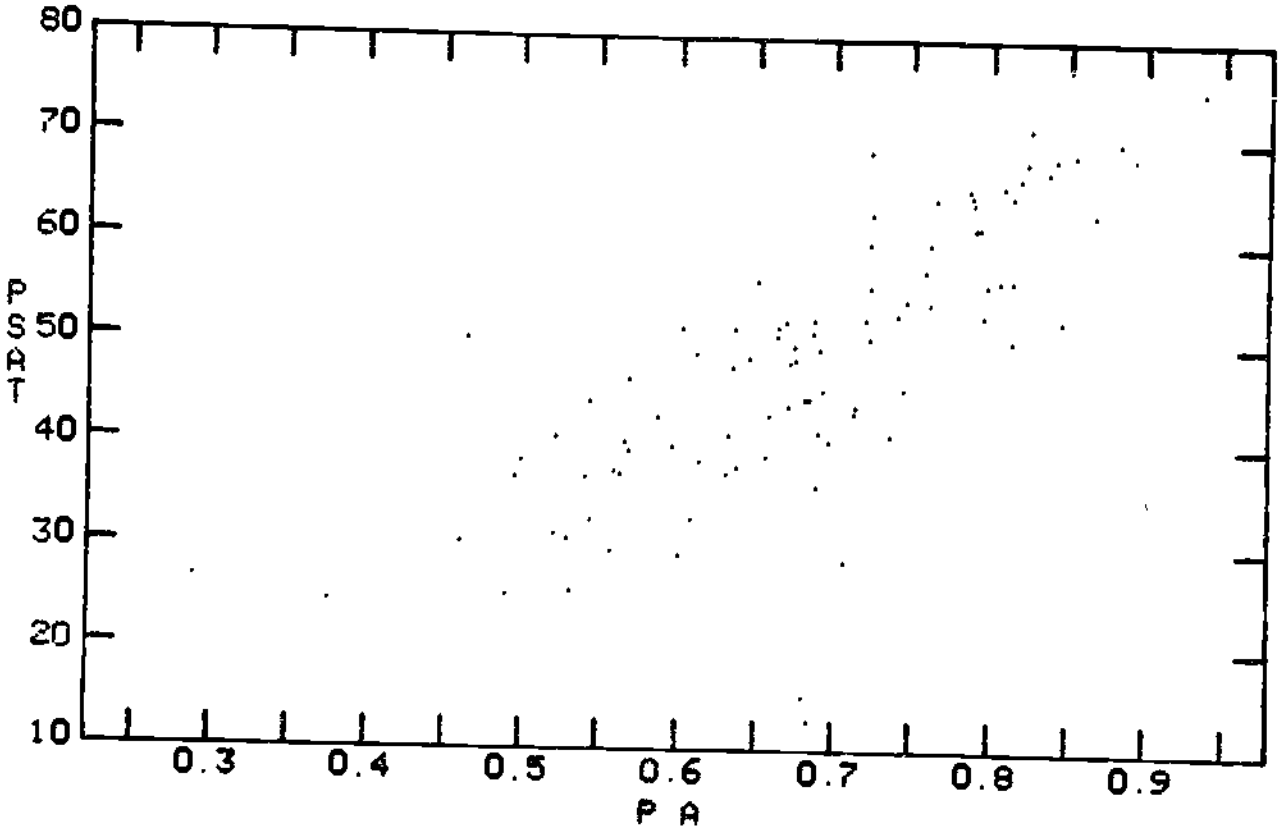


Figure 3.7bb PSAT vs. p (Form A)

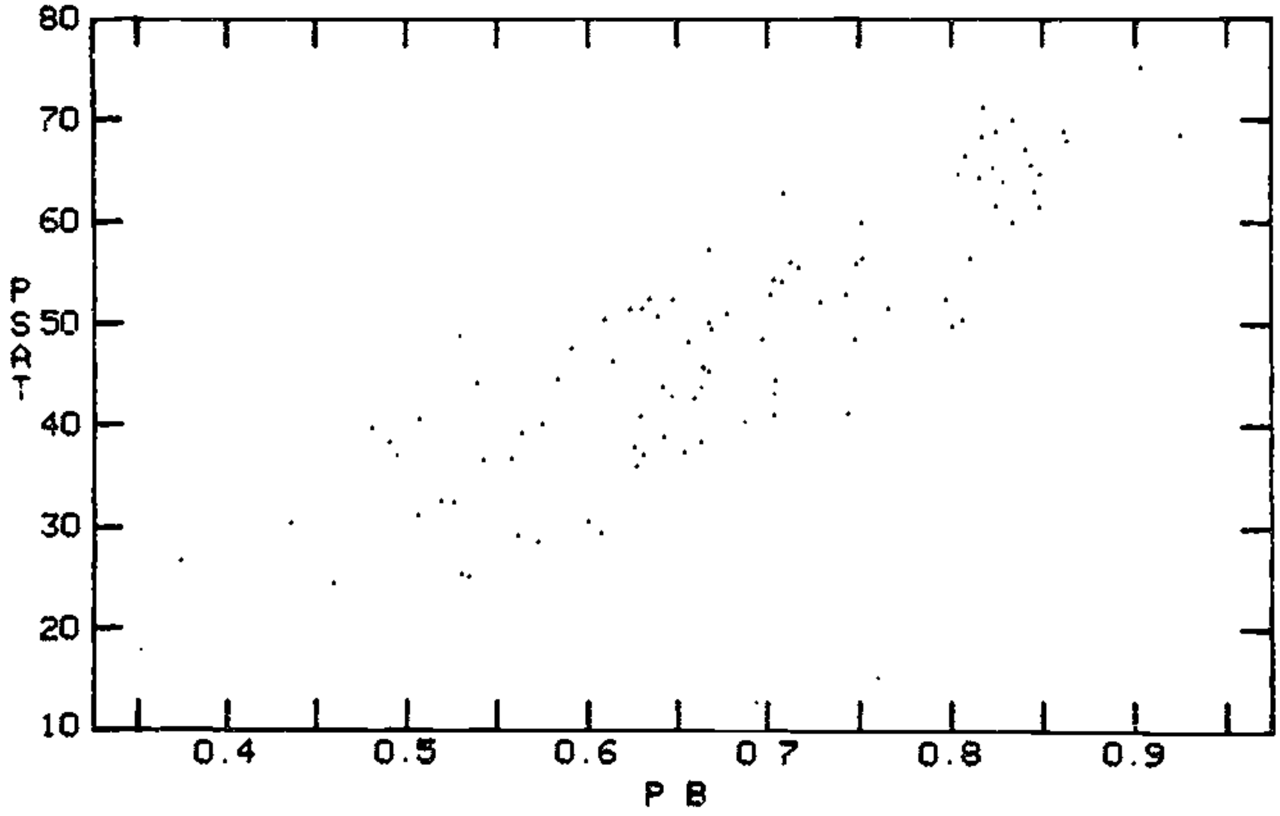


Figure 3.7cc PSAT vs. p (Form B)

Table 3.7d

Paired Tests Between Scores Calculated at the 25th Item for Forms A and Forms B of the BRTT (N=135)

Score	t-Test (P-Value)	One-Sample Test van der Waerden Test	(P-Value)
θ	2.0938 (.038164)	4.1465	($3.3756\text{X}10^{-5}$)
Ω	2.1517 (.033215)	4.2343	($2.2926\text{X}10^{-5}$)
b_{total}	2.0497 (.042341)	4.2791	($1.8761\text{x}10^{-4}$)
$b_{correct}$	1.8789 (.062434)	3.9204	($8.8402\text{x}10^{-5}$)
b_{high}	.34097 (.39438)	1.0353	(.30052)
b_{final}	3.4747 ($6.6947\text{X}10^{-4}$)	6.41	($1.4547\text{X}10^{-10}$)
$\underline{p}$	-.84969 (.39702)	-1.3169	(.18787)

Tests are computed on the difference, Form A minus Form B.

3.8 Parallellism and Reliability Across All Items

The previous section presented data on the reliability of the BRTT for scores computed at the 25th item. Here we consider the reliability of the test at all items.

Figure 3.8a shows the reliability of the BRTT as compared to that of the PSAT at 1, 2, ..., 25 items. The reliability for the BRTT was obtained by correlating the scores for each subject for Forms A and B at each step. The reliability of the PSAT was obtained bv adjusting the published reliability of the 65 item test to lengths of 1, 2, ..., 25 items by means of the Spearman-Brown formula.

Figures 3.8b through 3.8e show the comparison of the BRTT to the PSAT when the scores θ, $\underline{p}$, b_{total}, and $b_{correct}$ are used. As with θ, the reliability of the BRTT is higher at all points.

Figures 3.8f -- 3.8j show the mean value for the sample for each of the scores discussed above. In reading these graphs, note that the solid line represents Form A and the broken line represents Form B (in the previous graphs the solid line indicated the BRTT and the broken line the PSAT).

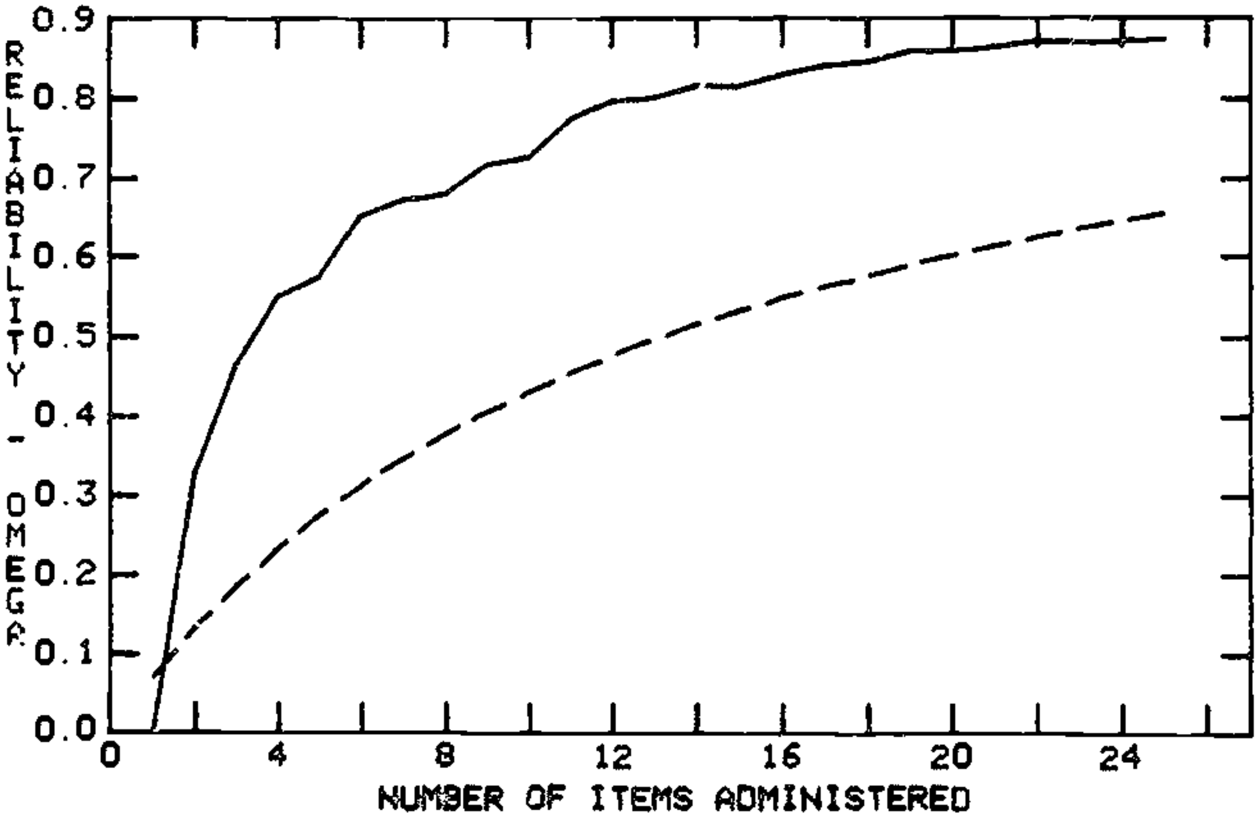


Figure 3.8b Reliability of Omega Compared to Reliability of PSAT.

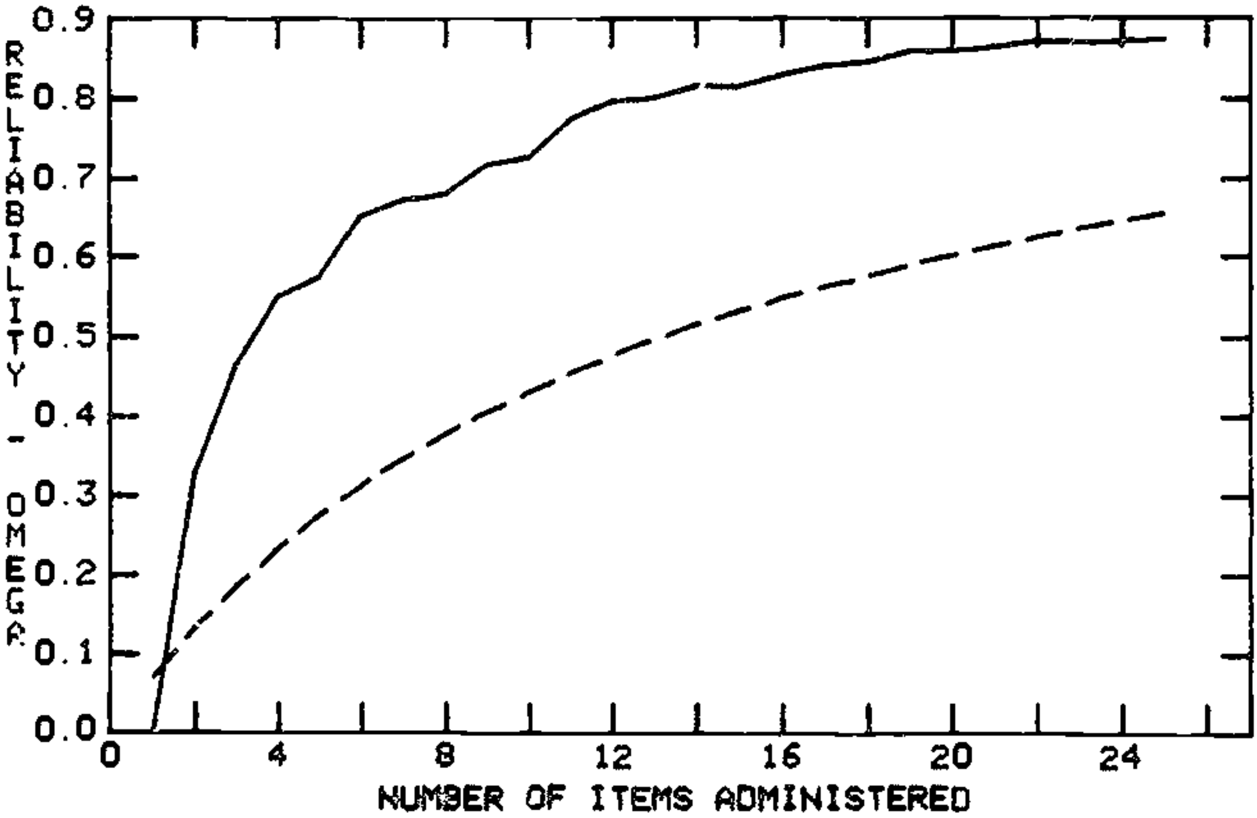


Figure 3.8b Reliability of Omega Compared to Reliability of PSAT.

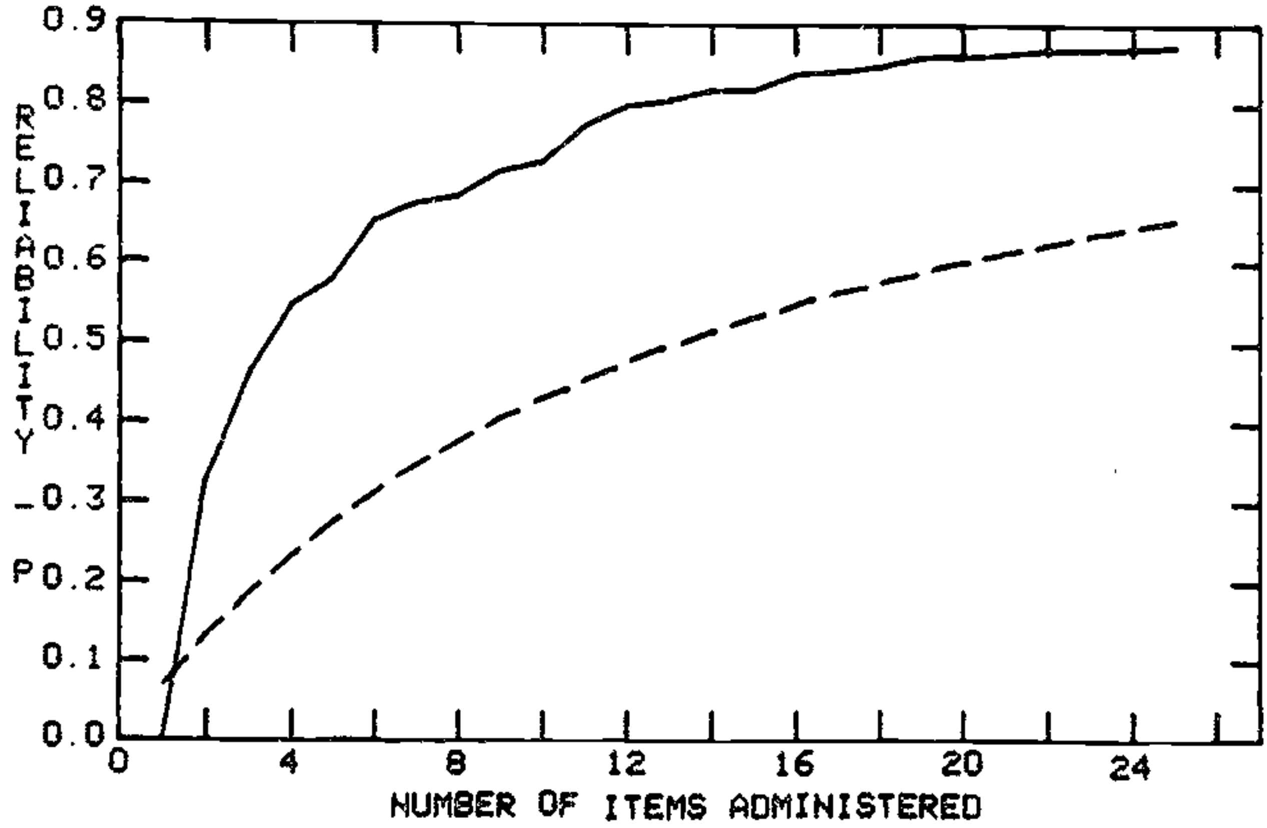


Figure 3.8c Reliability of p Compared to that of PSAT.

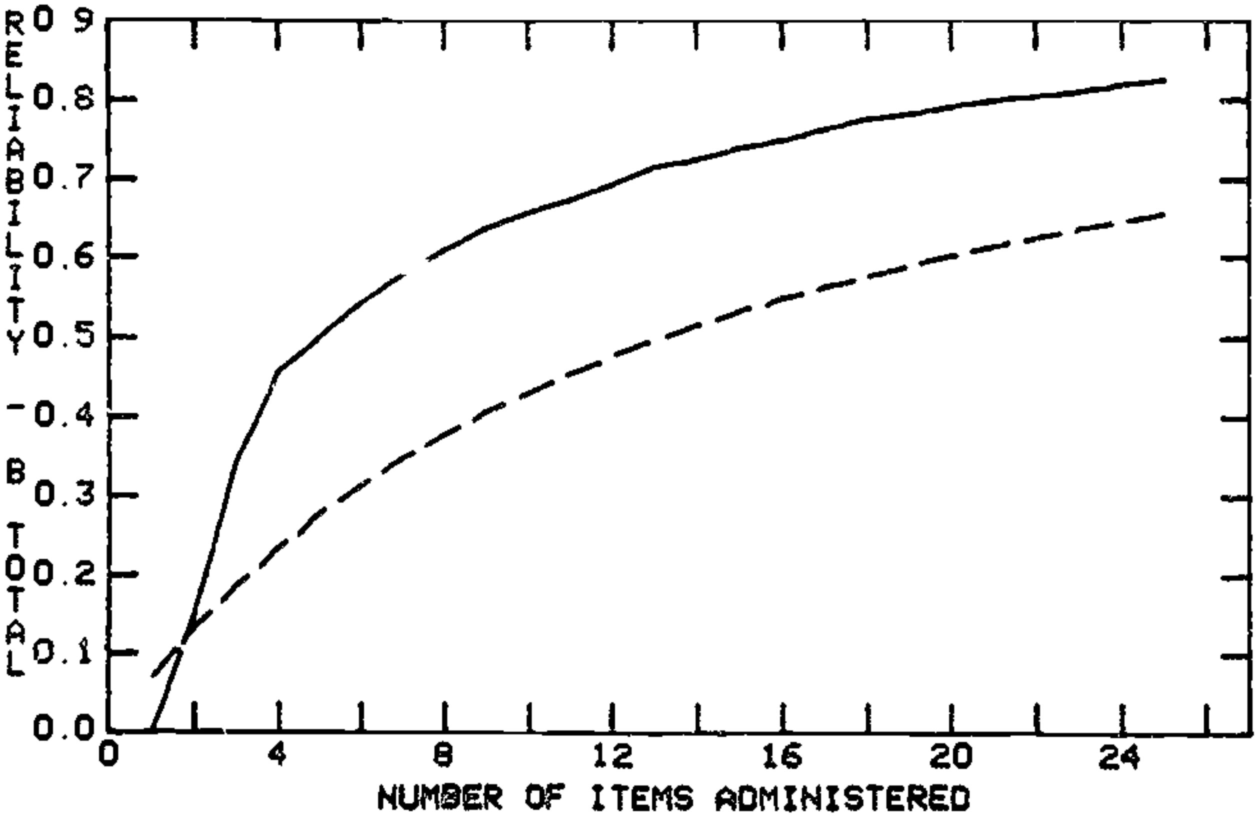


Figure 3.8d Reliability of b_{total} Compared with PSAT Reliability.

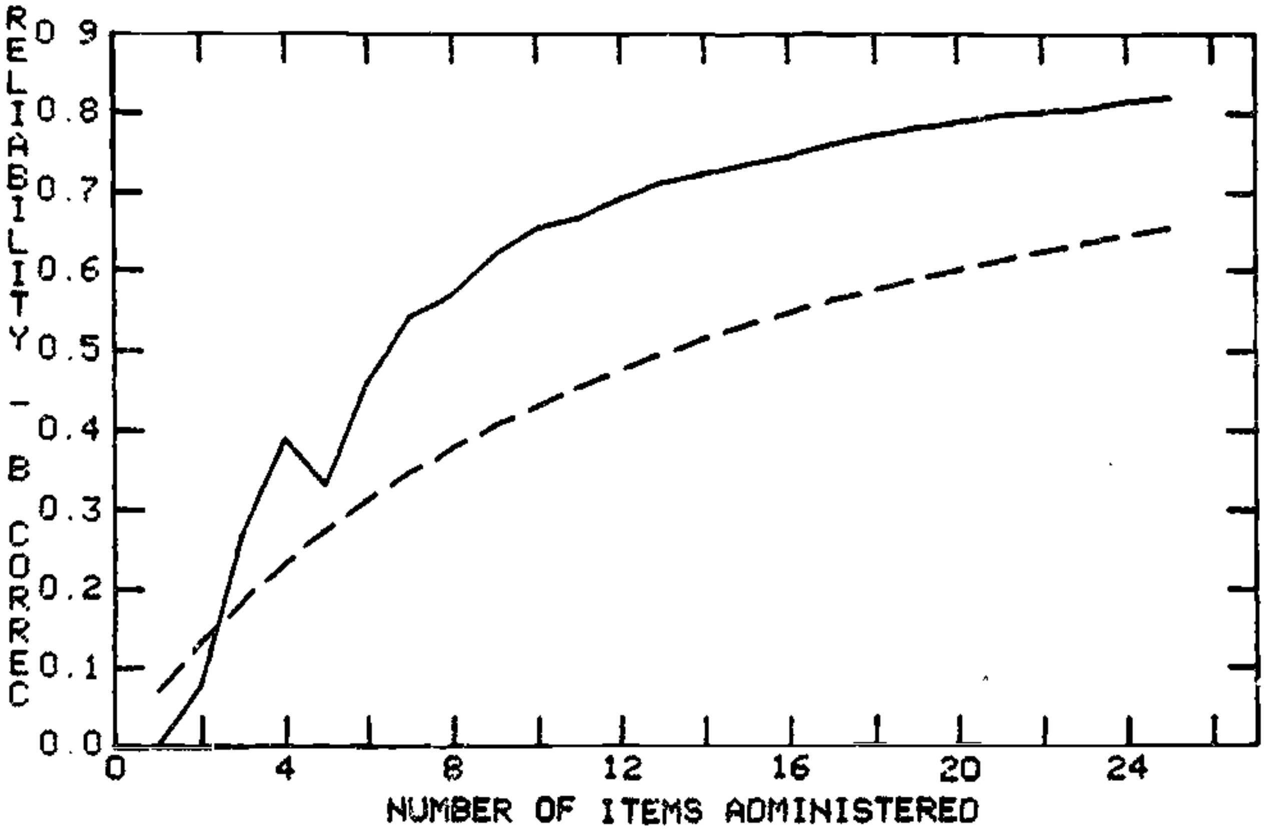


Figure 3.8e Reliability of $b_{correct}$ Compared with PSAT Reliability.

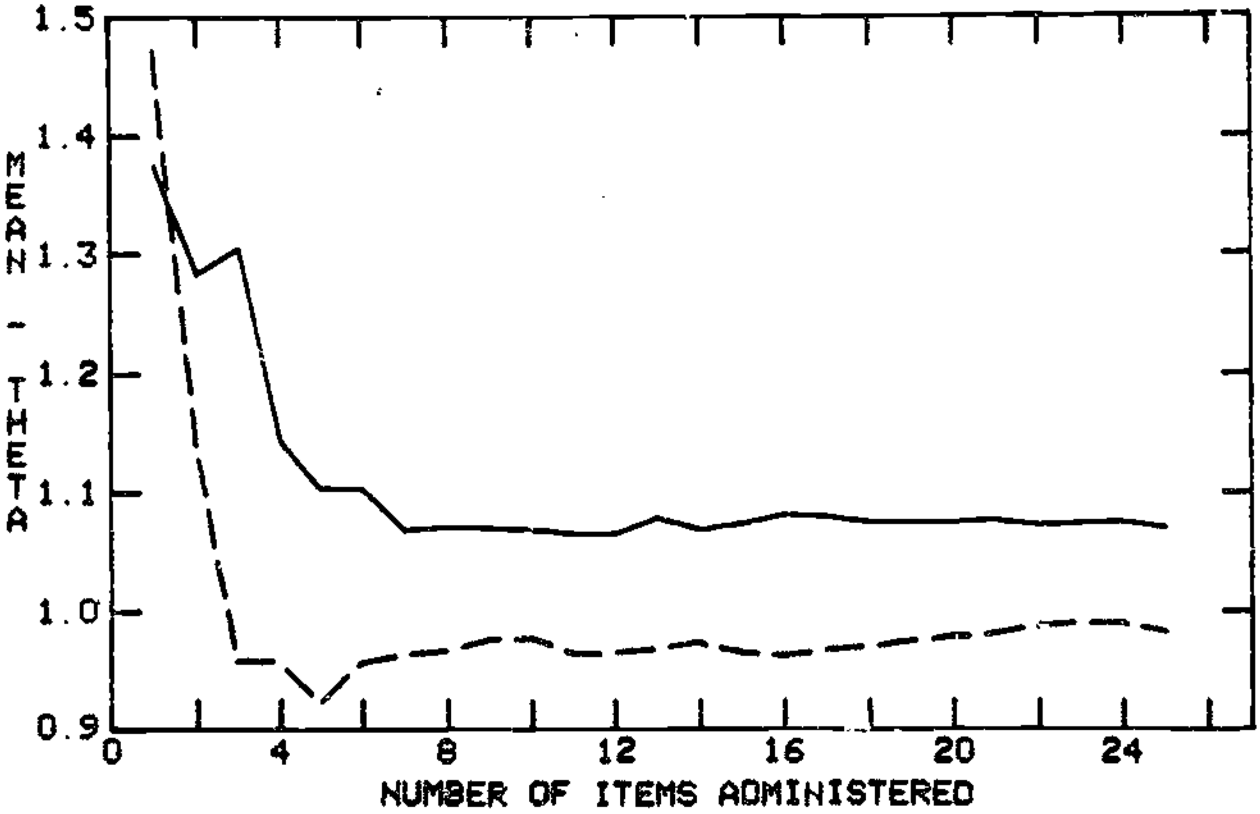


Figure 3.8f Mean of Sample for Theta

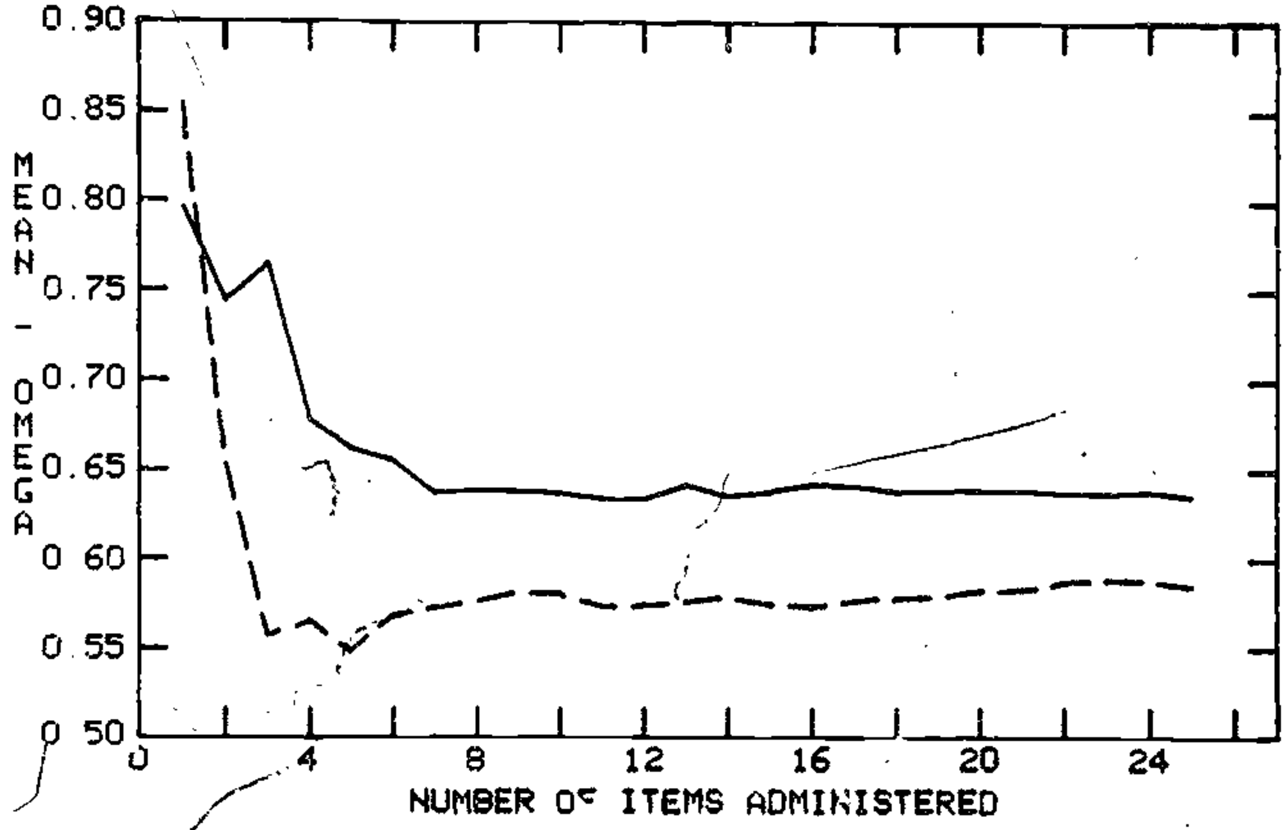


Figure 3.8g Mean of Sample for Omega

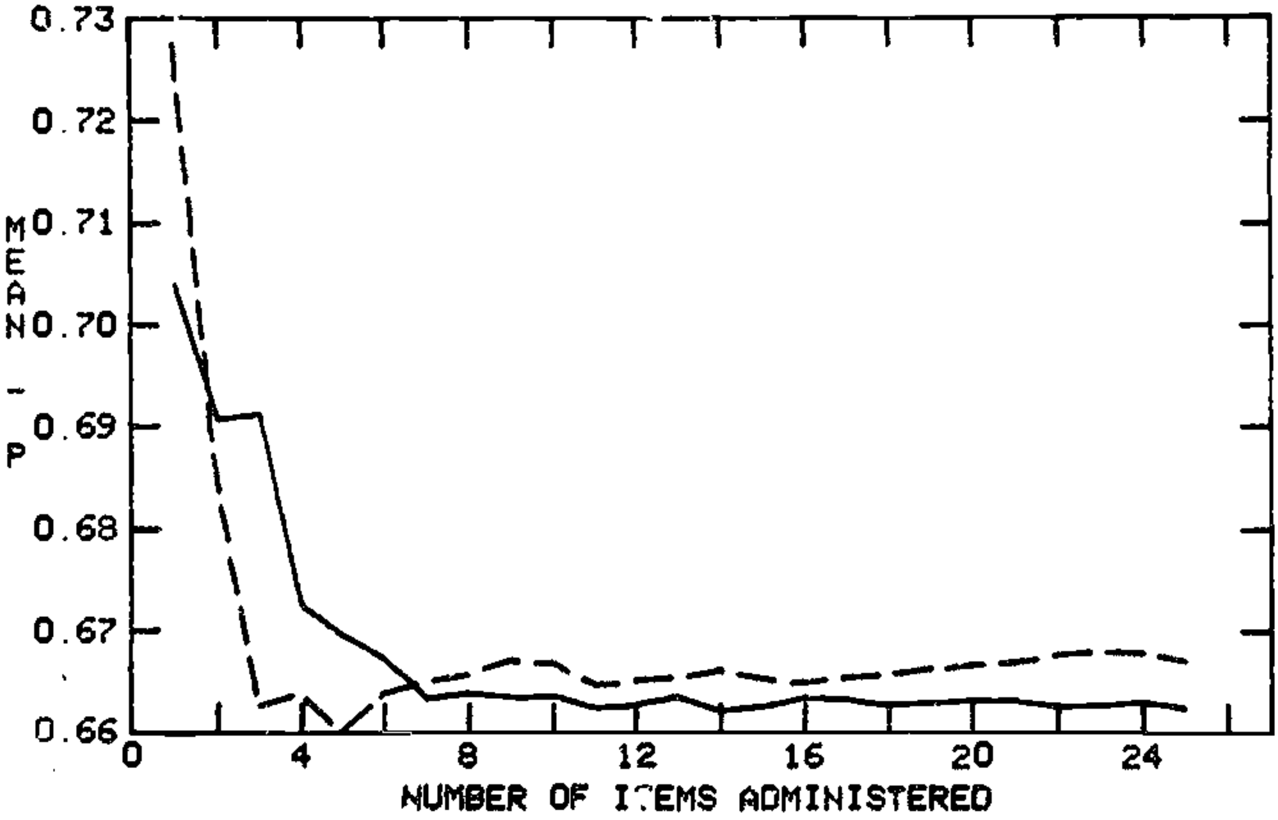


Figure 3.8h Mean of Sample for p

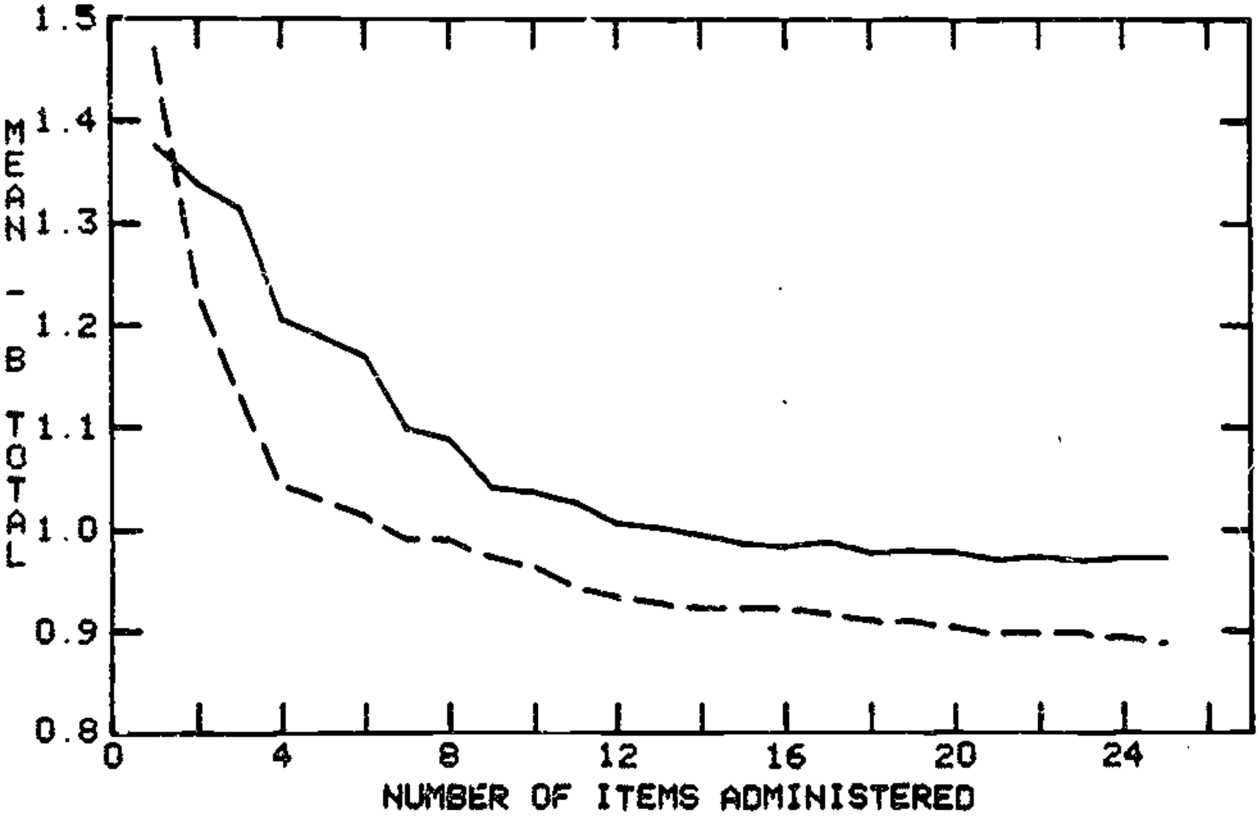


Figure 3.8i Mean of Sample for b_{total}

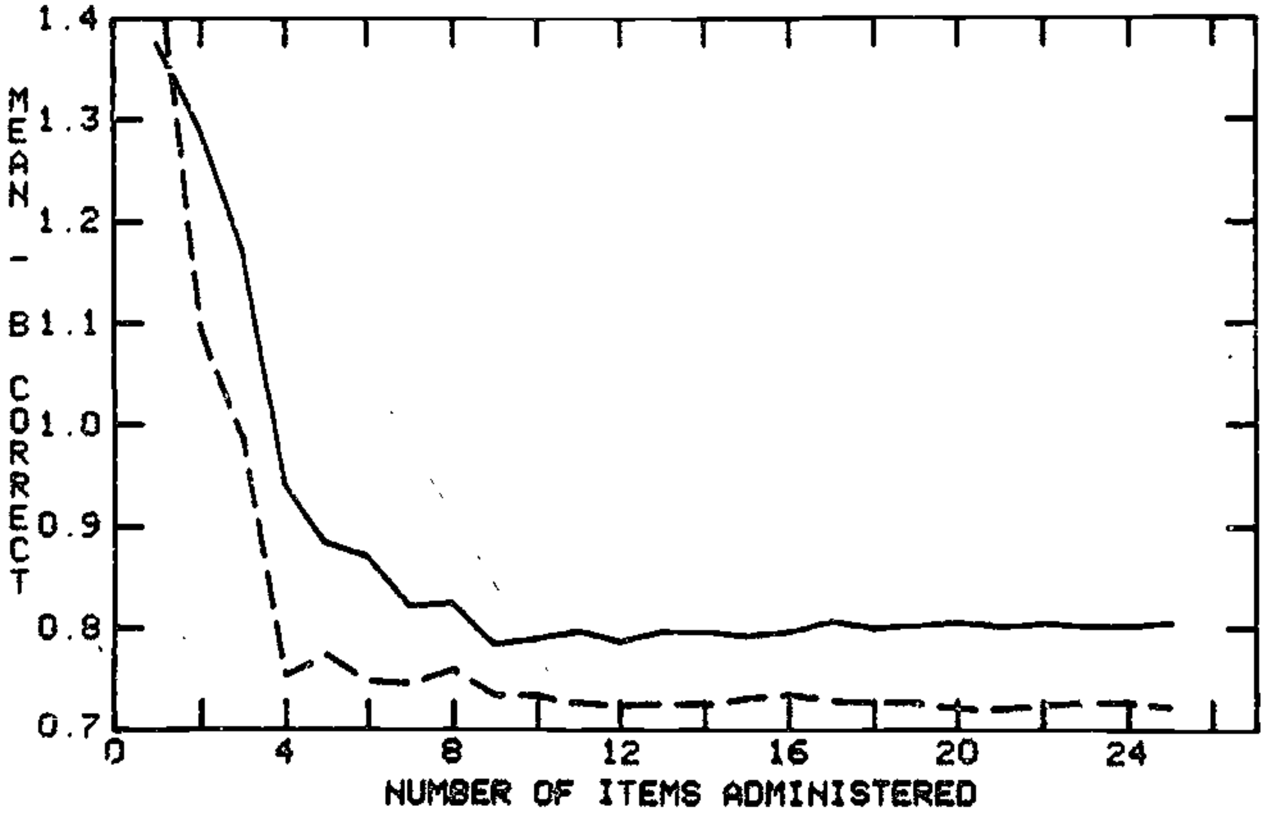


Figure 3.8j Sample Mean for $b_{correct}$

3.9 The Likelihood Function

The BRTT employs an item bank in which three parameters have been previously estimated for each item. These parameters are (a) discrimination (b) difficulty and (c) guessing. The parameters define an item characteristic curve which describes the probability of a correct response to the item given a trait level Θ. It is assumed that trait levels vary continuously and that the probability of a correct response to the item is an increasing monotone function of the trait level. Given trait level Θ, the item characteristic curves model the probability of observing a particular response vector. In the BRTT score, the problem is reversed: given an observed response vector, we estimate the trait level which gave rise to the observed pattern. The procedure used to score the BRTT is the maximum likelihood technique; the likelihood function is defined by the equation:

$$L(\theta) = \prod_{1}^{k} P_i(\theta)^{u_i}[\,1 - P_i(\theta)]^{1-u_i}$$

where $P_i(\theta) = P(u_i = 1 \mid \theta)$.

The desired estimate of Θ is the point at which the likelihood function is maximized. In general, the likelihood function assumes the form shown in Figure 3.9a. However, if the examinee answers all items correctly, the function becomes asymtotic to +1 and assumes the form shown in Figure 3.9b. In this case, the maximum is taken at + ∞, resulting in an estimate for Θ equal to + ∞. Similarly, if the examinee answers every question incorrectly, the likelihood function will assume the form shown in Figure 3.9c and will yield an MLE of −∞. Therefore, the MLE estimate of Θ can only be employed if the examinee's responses produce a likelihood function with maximums obtained at a realvalued number. Given that the estimate of Θ is non-infinite, the likelihood function will tend to provide increasingly precise estimates as the length of the response vector increases.

Figure 3.9d shows the likelihood functions observed for a single individual taking Form A of the BRTT. The item difficulty ad'iinistered and the MLE of Θ is shown at each step. Note that the likel hood function assumes infinite values at steps 1 and 2; the estimate of Θ at these points is set equal to the difficulty of the item admi istered. The estimates become stable fairly rapidly.

Figure 3.9e shows the analogous graphs for the same individual on Form B of the BRTT.

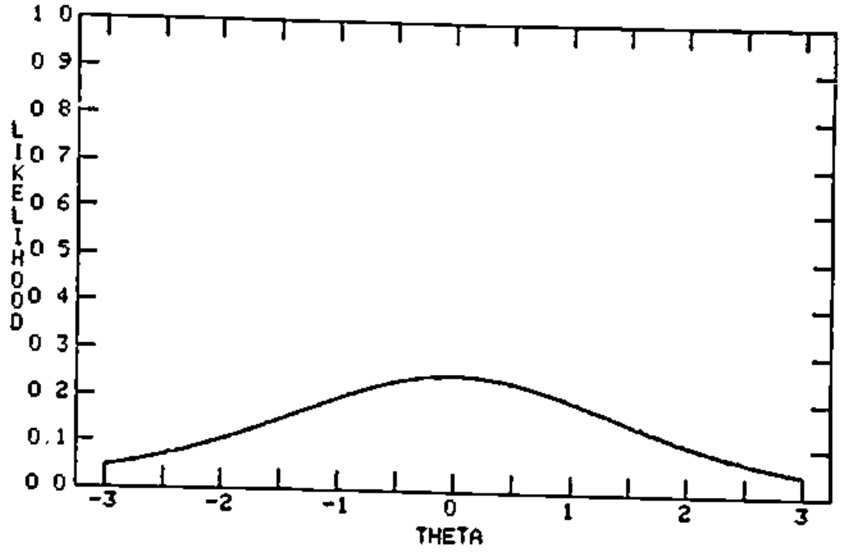


Figure 3.9a A Typical Likelihood Function

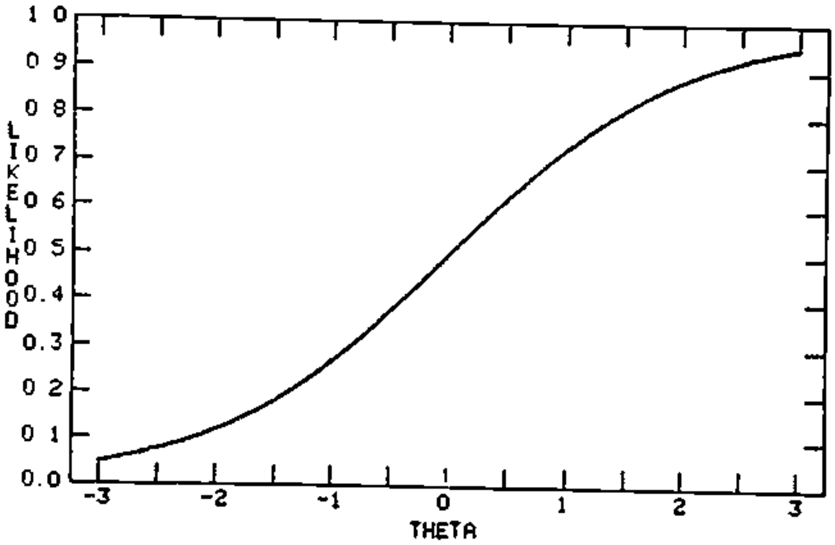


Figure 3.9b A Likelihood Function with Maximum +∞

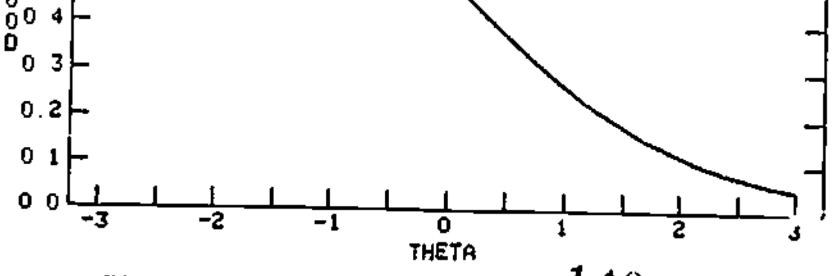


Figure 3.9c A Likelihood Function with Maximum -∞

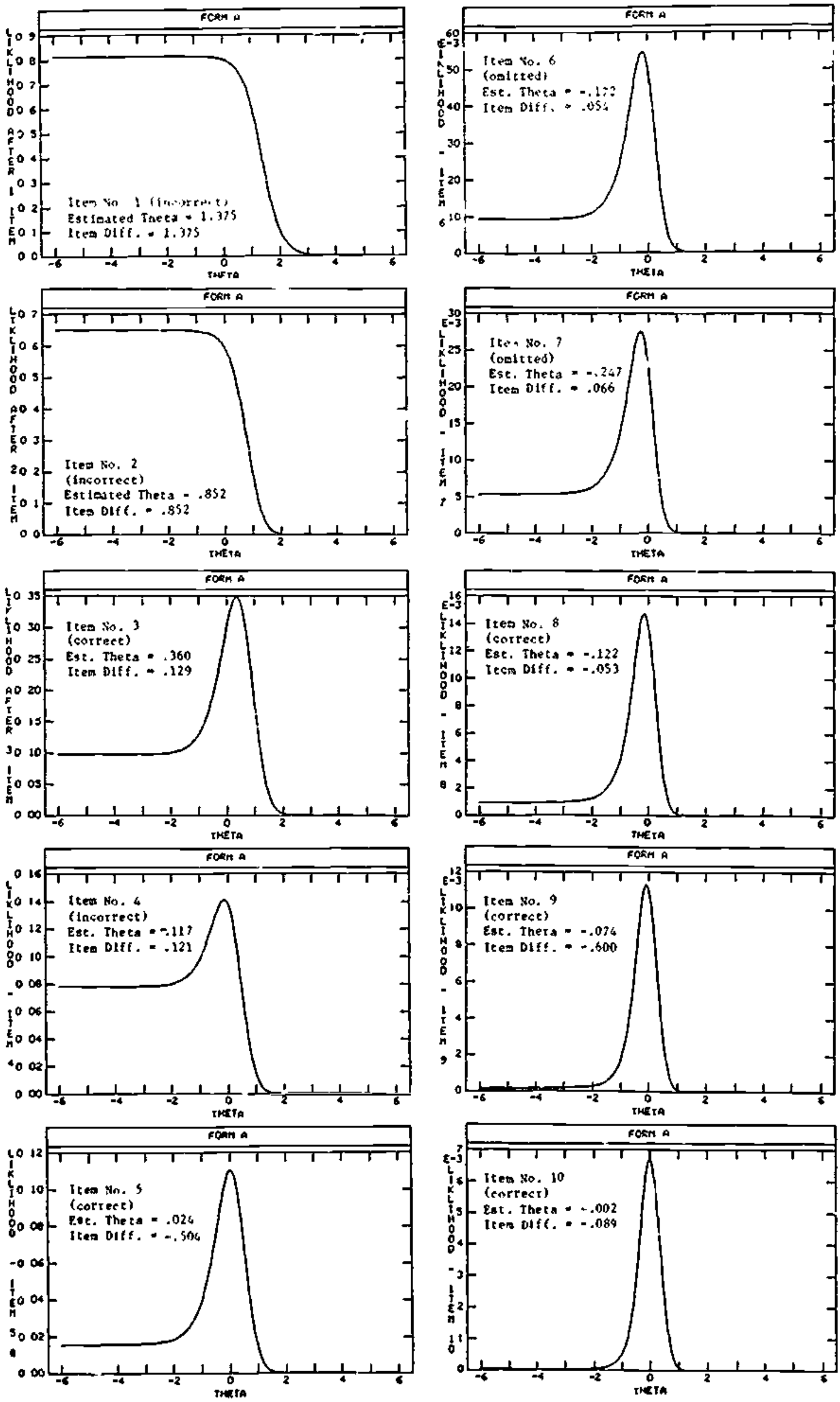


Figure 3.9d Observed Likelihood Function for a Single Student (Form A)

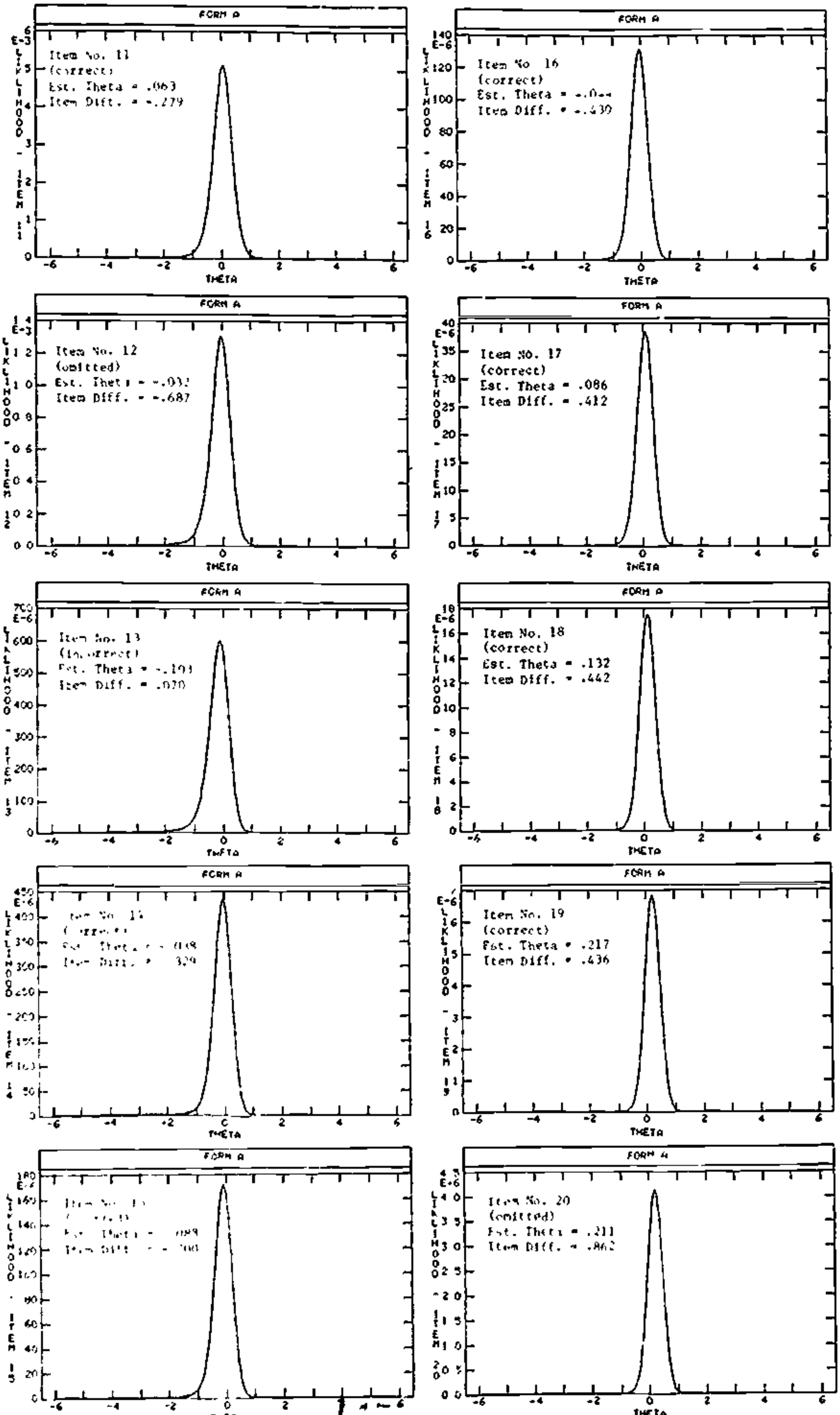


Figure 3.9d continued

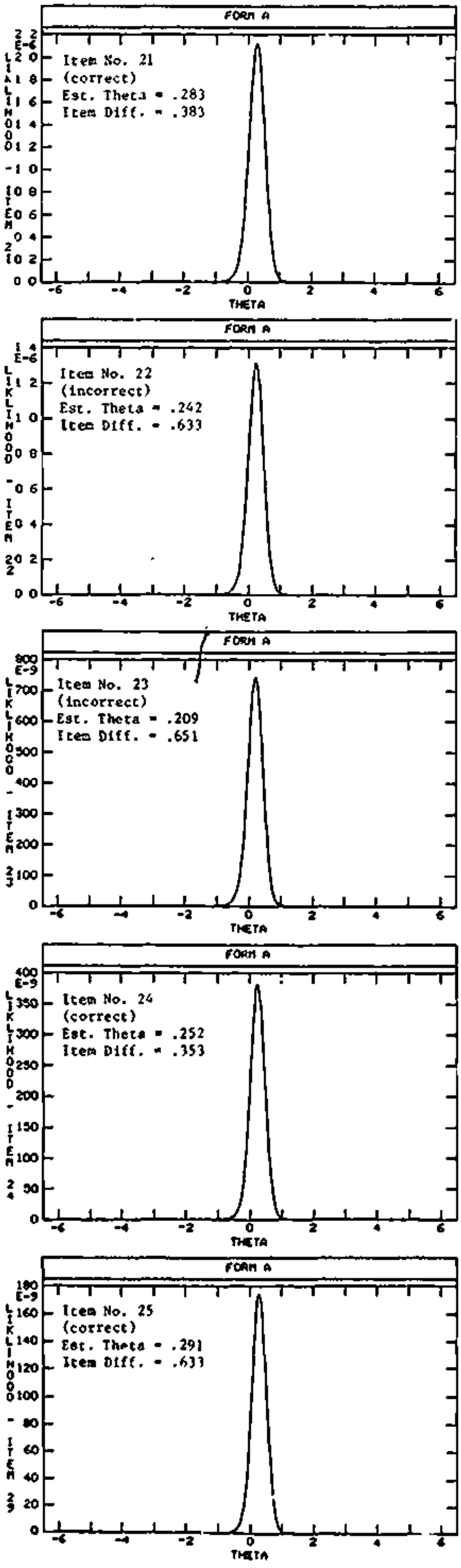


Figure 3.9d continued

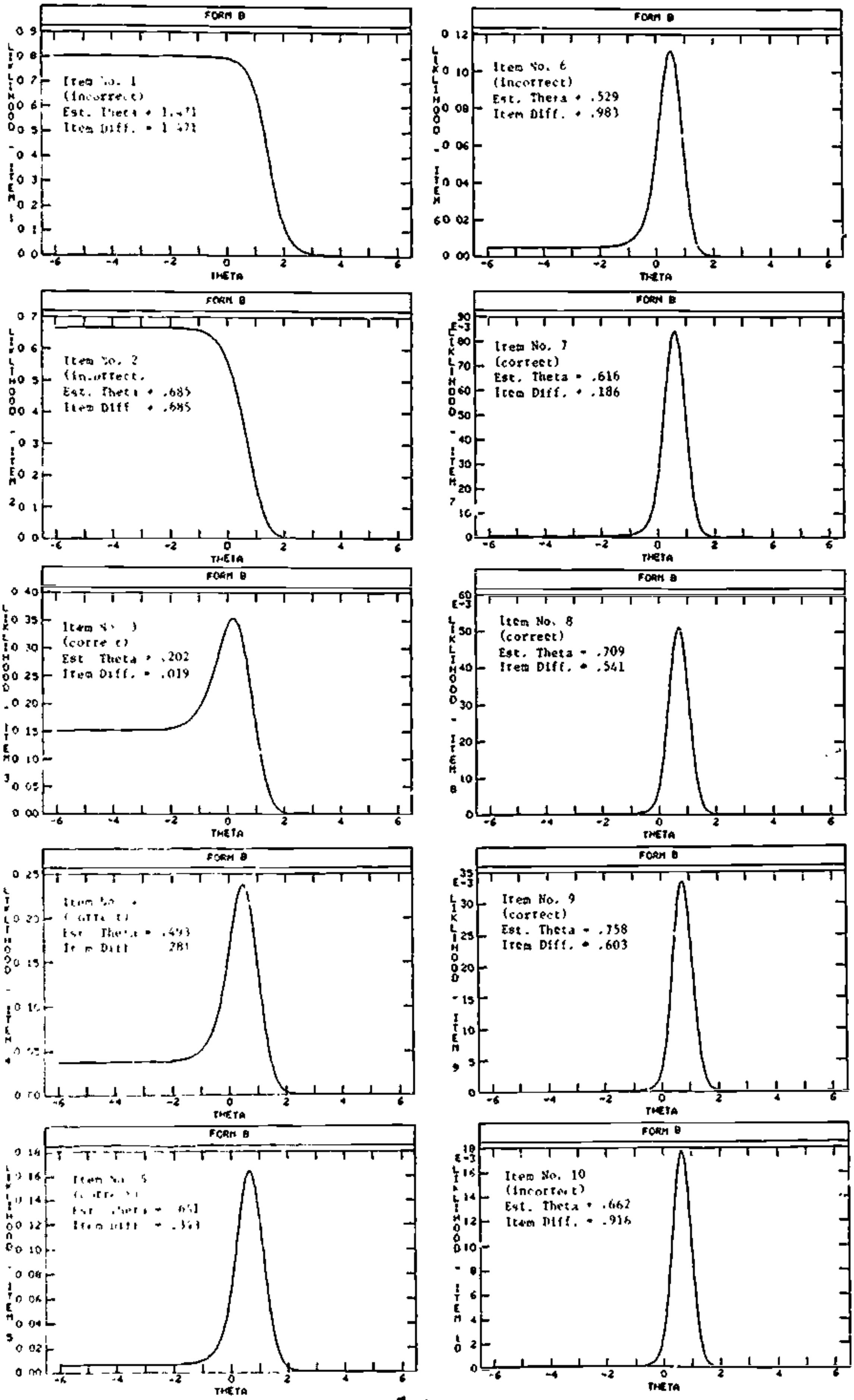


Figure 3.9e Likelihood Functions for a Single Student (Form B)

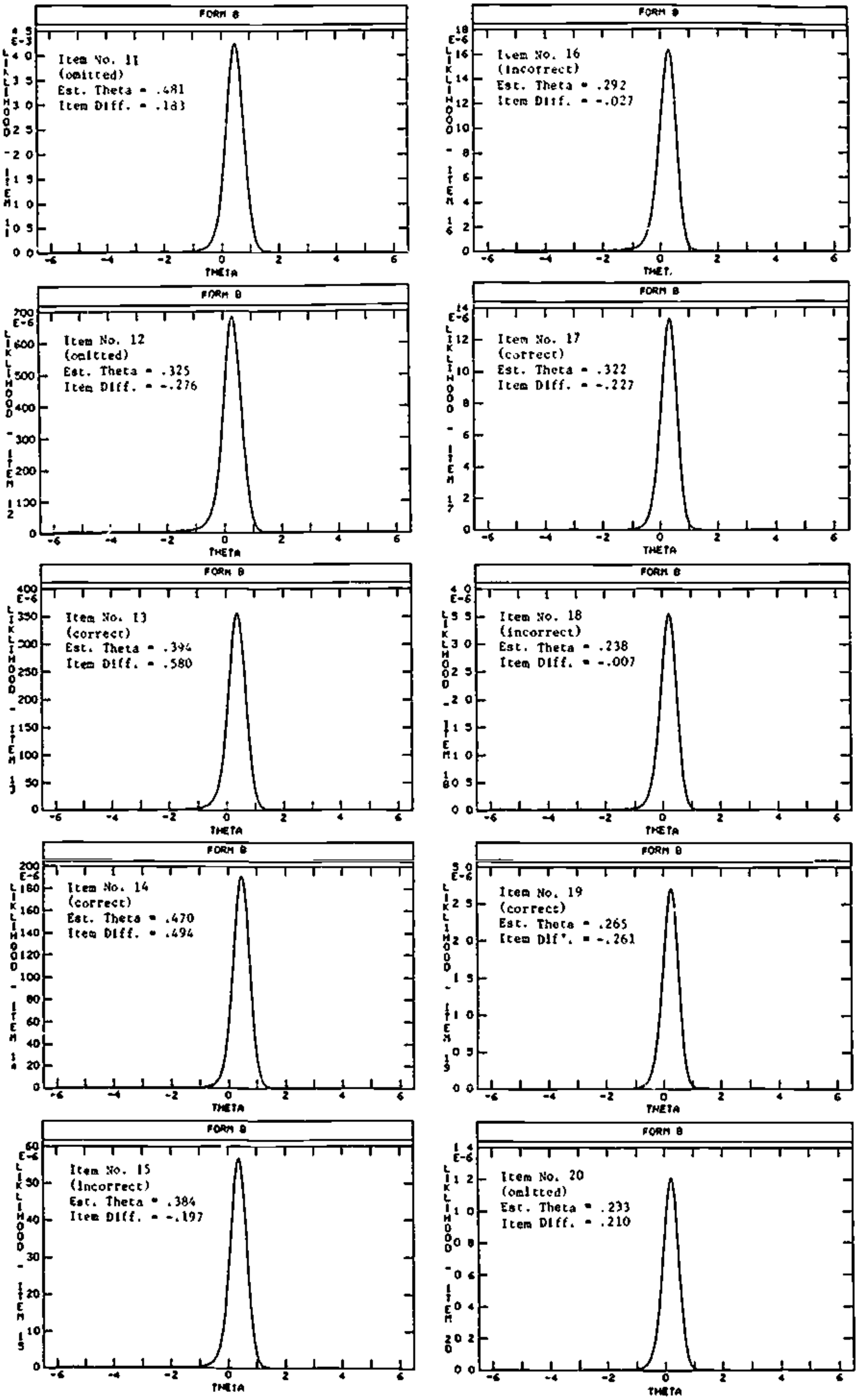


Figure 3.9e continued

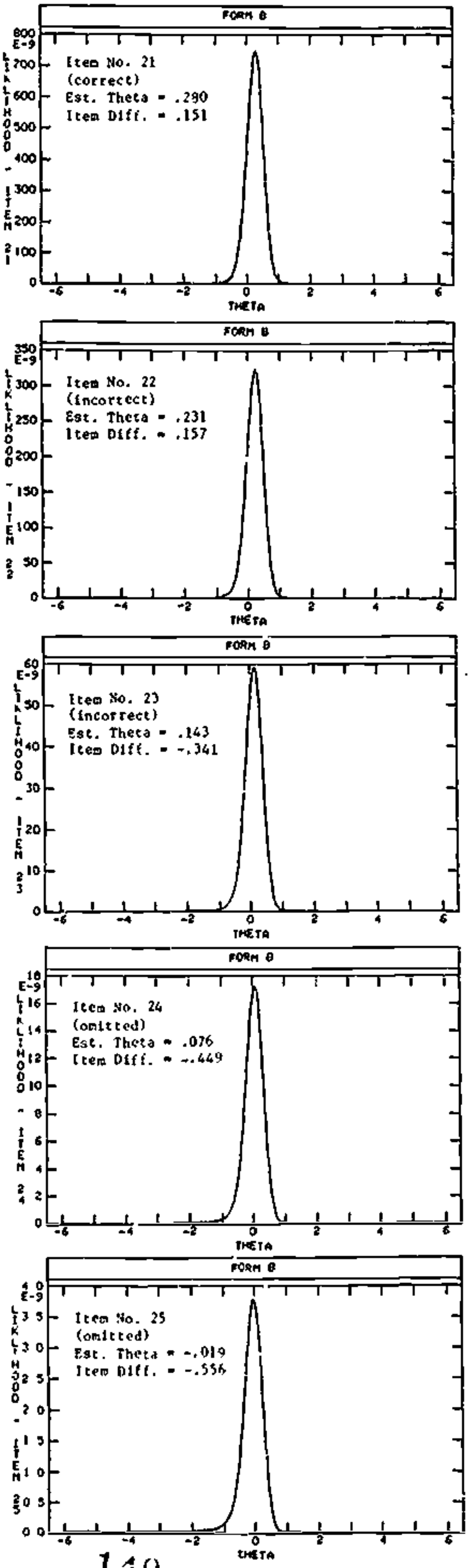


Figure 3.9e continued

3.10 Analysis of the Numerical Algorithm for Determining θ

We investigated the performance of the system with respect to the computation of the maximum likelihood estimation. Several interesting results were obtained which suggest refinements to future operational systems.

As described in section 3.9, the CAT system uses the statistical method of maximum likelihood for estimating an individual's ability parameter given his string of responses to the items chosen by the computer. The numerical technique employed for determining this ability parameter is the Newton-Raphson method. The exact mathematical details will be deferred for the moment so that we may focus on two functions of theta which depend on the examinee's string of responses. Given the ability parameter θ, denoting the probability of correct response to the ith item by $P_i(\theta)$ and the derivative with respect to θ of this fuction by $P_i(\theta)$; define the weight for item i as

$$(1) \qquad w_i(\theta) = \dot{P}_i(\theta)/[P_i(\theta)\ Q_i(\theta)]$$

where $Q_i(\theta)$ is the probability of incorrect response given $\theta = 1-P_i(\theta)$.

Defining $u_i = 1$ for a correct response and $u_i = 0$ for an incorrect response to item i, we have

$$(2) \qquad S(\theta) = \Sigma u_i w_i(\theta).$$

This is the weighted number right score described in Section 3.3.

The second function is the expected weighted number correct given θ and is defined as

$$(3) \qquad T(\theta) = E\{S(\theta)|\theta\}.$$

The function can be further specified by using the relationship from item response theory that states that the expected value of the response u_i to item i is $P_i(\theta)$; hence,

(4) $$T(\Theta) = \Sigma P_i(\Theta)\, w_i(\Theta)\ .$$

The statistical theory of maximum likelihood estimation reduces in this case to finding the value of Θ that satisfies the equation

(5) $$S(\Theta) = T(\Theta)\ .$$

In other words, the maximum likelihood estimation of the ability is that value which best fits the subject's responses in the sense that the weighted number right score is equal to its theoretical expected value.

Figures 3.10b and 3.10c are plots of the obtained S and T values for each examinee from Forms A and B of the BRTT. These figures show equality between S and T. The figures verify that the computer program in the CAT system is accurately finding the maximum likelihood estimator.

The maximum likelihood estimator of Θ is the value which maximizes the likelihood given the response vector $u = (u_1, \ldots, u_{25})$

(6) $$L(\Theta) = \Pi\ P_i(\Theta)^{u_i}\,[1 - P_i(\Theta)]^{1-u_i}$$

The value of Θ that maximizes $L(\Theta)$ can also be obtained as the maximum of the logarithm of the likelihood:

(7) $$l(\Theta) = \Sigma\ u_i \log P_i(\Theta) + \Sigma(1-u_i)\log(1 - P_i(\Theta)$$

It can be verified that the derivative of $l(\Theta)$ with respect to Θ is

(8) $$\dot{l}(\Theta) = \Sigma u_i w_i(\Theta) - \Sigma P_i(\Theta)\, w_i(\Theta)\ .$$

From definitions of $S(\Theta)$ and $T(\Theta)$ in equations (2) and (4), we see that this is equivalent to

$$\dot{1}(\Theta) = S(\Theta) - T(\Theta) \qquad (9)$$

So that the maximum likelihood estimator which satisfies equation (5) equivalently satisfies the equation

$$\dot{1}(\Theta) = 0.$$

This is true since the extrema of a function, in particular $1(\Theta)$, may be found, under suitable assumption, by setting the derivative or rate of change equal to 0. This fact follows from observing that in Figure 3.10a the line tangent to $1(\Theta)$ at Θ has slope equal $\dot{1}(\Theta)$ and that the maximum of Θ is obtained when the line is flat, that is, when the slope is 0.

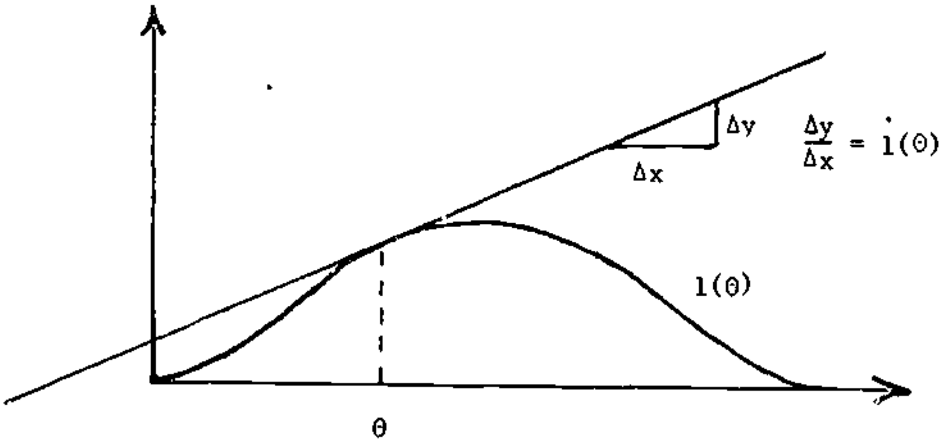


Figure 3.10a

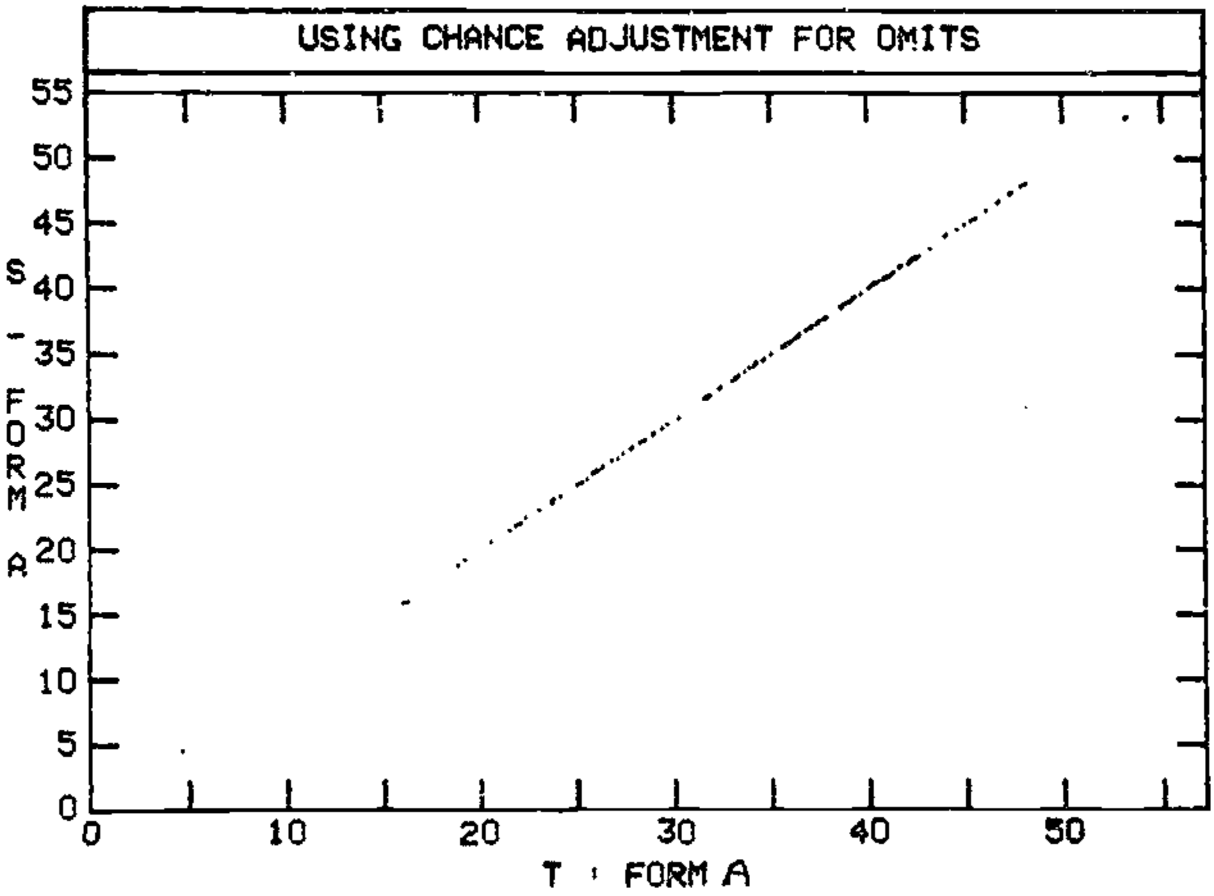


Figure 3.10b Relationship Between Observed Weighted Number Correct to Expected Number Correct (Form A) Evaluated at the Reported MLE of Theta.

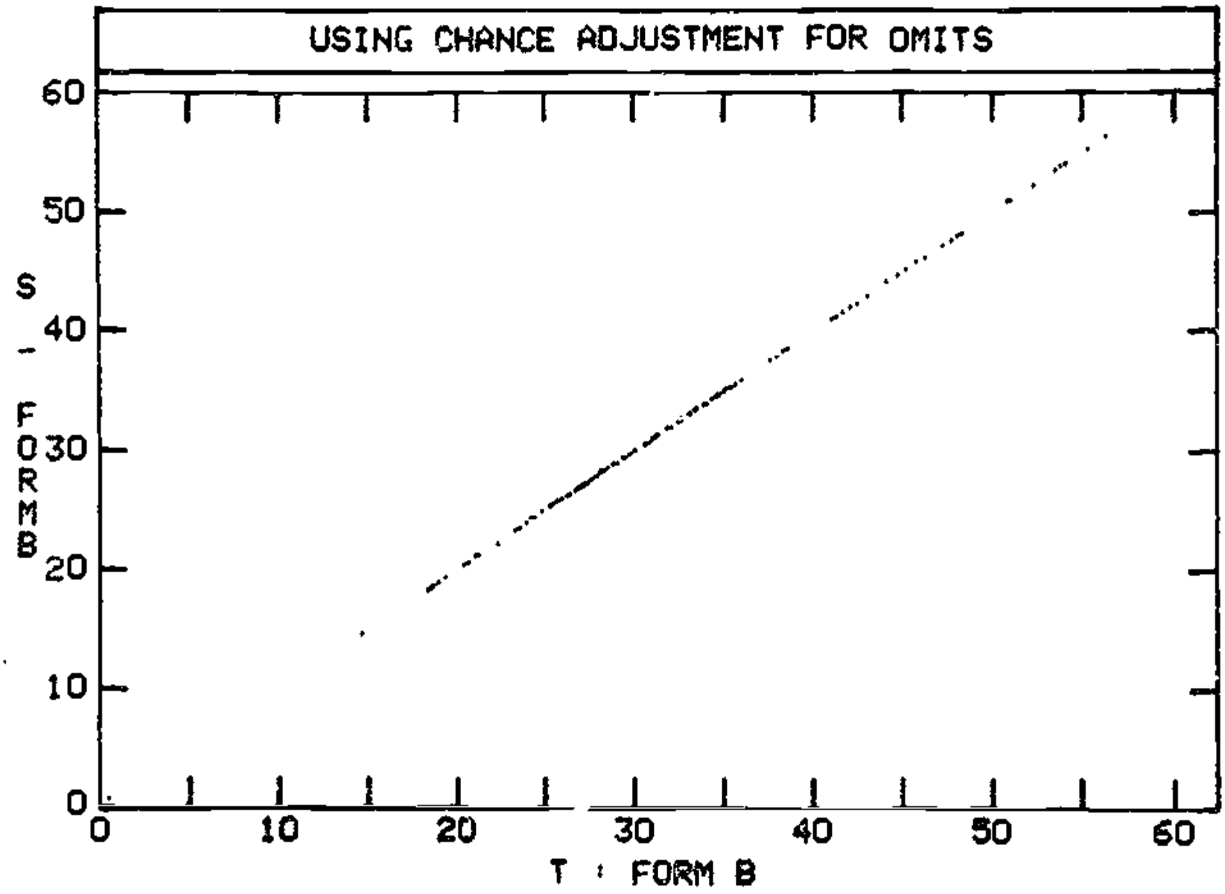


Figure 3.10c Relationship Between Observed Weighted Number Correct to Expected Number Correct Evaluated at the Reported MLE of Theta.

3.11 Monte Carlo Analysis of the Item Selection Procedure

The BRTT chooses items for an individual examinee which best suit his or her ability. In the ideal situation, items are chosen so that the examinee has about a 65% estimated chance of responding correctly and the responses to such items give the most information about the examinee's ability. The BRTT would use prior information on the individual to choose all the items to match the examinee's ability so that each response would be maximally informative. However, prior to administration of the first item no specific information about the individual's ability exists. As responses are accumulated, the BRTT obtains progressively better estimates of the examinee's ability so that estimates toward the end of the test are more accurate than estimates at the beginning. This process was illustrated graphically in section 3.9. Because the precision of estimation changes over the course of the test, the BRTT will administer some items which are too difficult and some which are too easy for the particular examinee.

A rough idea of how far the actual item selection procedure deviates from the ideal case may be obtained by examining the relationship between the number of correct responses and the final estimate of ability. Graphs 3.11a and 3.11c display the number correct vs θ, the measure of ability, for Forms A and B. Under the ideal circumstance of having the BRTT administer items such that the probability of a correct response is .65: no regression of number correct on the estimate of ability would be expected (assuming that the final estimate is very close to the examinee's true ability). However, the graphs reveal some regression -- particularly for Form A. The correlations between number correct and theta are .6970 for Form A and .5616 for Form B.

One way of studying the correspondence of the assumed model for response to the true underlying model is to compare the examinee's actual responses to the responses obtained by simulating each examinee's responses based on the estimated θ. For each item actually administered to an examinee having the estimated theta, the simulated response was obtained by randomly generating a number in the unit interval. If the number was greater than P(θ), the response was taken to be incorrect; if it was less than P(θ), the response was taken to be correct. If the item response probabilities are modeled accurately and the estimated θ is reasonably accurate, the same pattern should appear in the simulated number correct as in the observed number correct. Graphs 3.11b and 3.11d display the simulations for Forms A and B. The patterns are not very similar; the relationship between the simulated and live data is not the same.

The findings of the Monte Carlo study suggest the need for additional research into the determinants of individuals' responses to items.

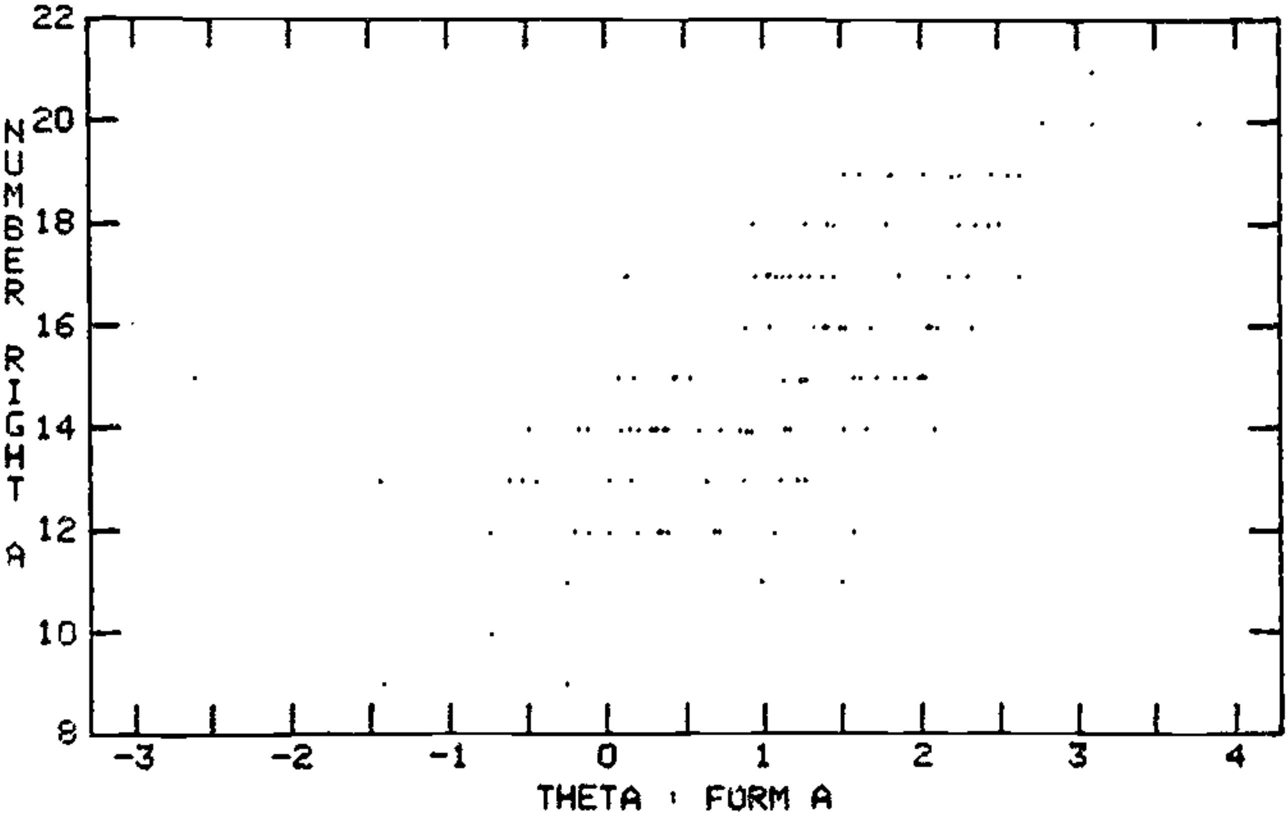


Figure 3.11a Observed Data - Form A

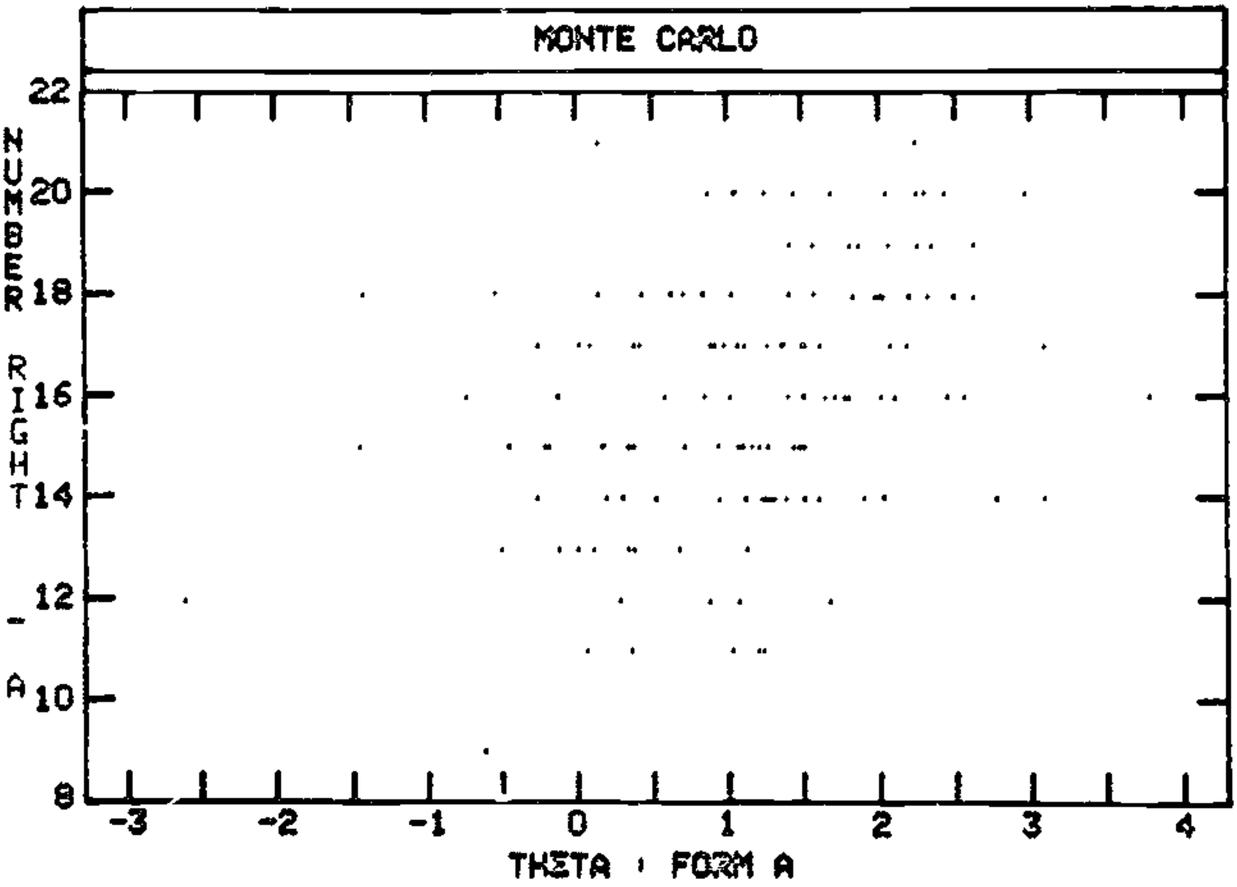


Figure 3.11b Simulated Data - Form A

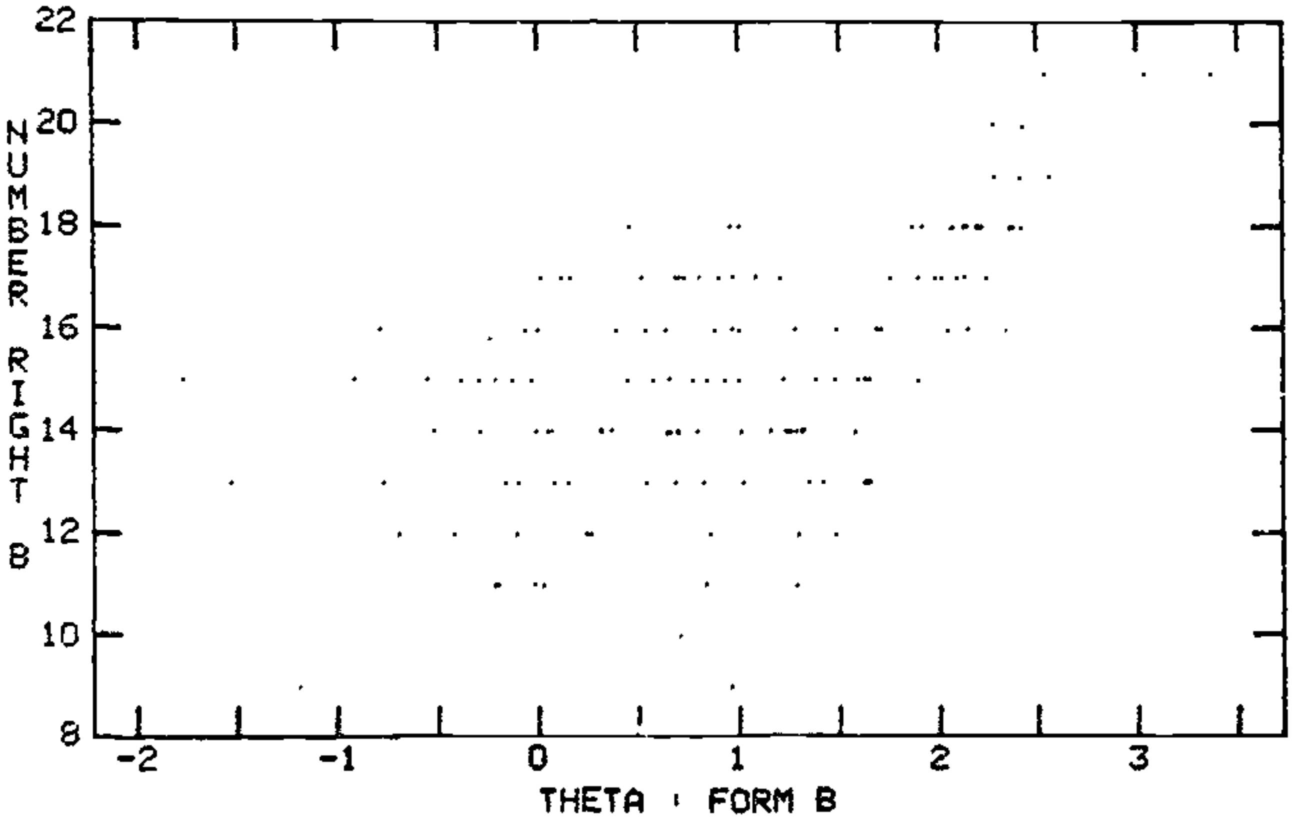


Figure 3.11c Observed Data - Form B

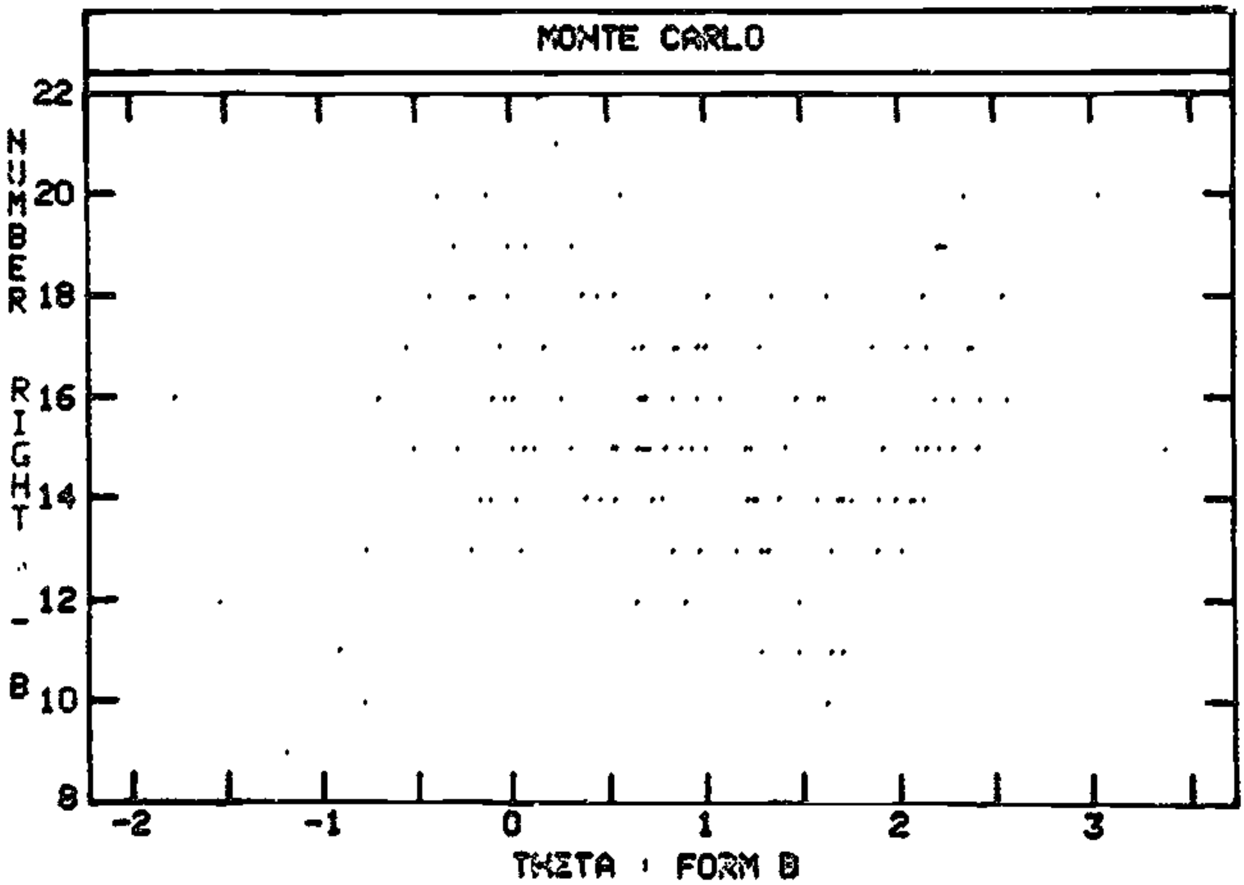


Figure 3.11d Simulated Data - Form B

3.12 Anomalous Cases

This section presents all cases which the BRTT system failed to process optimally, and it is hoped that this data can be used for improvements in the system.

As stated previously, eleven cases were eliminated from the analyses: six because the BRTT system reported a final Θ of -6 on at least one of the forms, three because the Form A to Form B final estimates of Θ were inconsistent, and two because their Form A estimates of Θ were $+\infty$ after the 8th item and $-\infty$ after the 9th item. The second of these three cases was caused by loss of attention or fatigue on the second test administered. The remaining cases were anomalous due to a feature of the type of numerical algorithm used by the BRTT to determine the maximum likelihood estimator.

Martha Stocking of ETS found that a modified Newton-Ralphson procedure could be successful in overcoming the infinite Θ problem in seven of the cases. In the remaining case, the Θ of -6 was a proper approximation to the maximum likelihood estimator since the likelihood function, due to the examinee's responses, attained the maximum at $-\infty$. The transformation of $\Theta = -6$ to the p score -- the expected proportion correct for the entire item pool -- is equal to the proportion correct if the examinee had no knowledge of the material and was guessing. This is a reasonable way to handle actual maximum likelihood estimates that equal $-\infty$.

Section 3.13 Summary and Conclusions

This section summarizes the data presented in Chapter III and highlights a number of significant findings.

Data log files created by the computerized test administration system were subjected to statistical analysis. Nine score variables were derived from the data and examined. Three scores of particular interest are θ, the parameter commonly used to denote the latent trait; Ω, a monotone transformation of theta; and p, the expected proportion of items correct over the entire item pool.

Frequency distributions of the scores revealed that both Form A and Form B had roughly comparable distributions. Tests for the normality of the sample data revealed some deviations from normality. It was decided not to attempt to transform the data to a more normal form since it was unlikely that the transformation function would generalize to subsequent samples. In some analyses, non-parametric statistics were employed to compensate for the lack of an underlying normal distribution.

A major finding of the study was that the information yield of the BRTT closely approximated the simulation results reported by Lord (1977a). This result was important because it confirmed that the accuracy of the BRTT conformed to theoretical expectation. Although this result must be interpreted with some caution, since the empirically observed information functions were calculated by use of the three parameter logistic model, the finding is broadly supportive of the utility of the three parameter model, in general, and the design of the BRTT. Monte Carlo simulations disclosed some discrepancies between theoretical expectations and observation. The simulation results suggest that there is need for a

3.13 continued

better methodology for finding mathematical models that adequately describe a subject's responses to individual items.

A second major finding of the study was that the BRTT proved to be highly reliable. The reliability of the 25 item test was .8719 which compares favorably with the 65 item PSAT whose reliability was .9111. Since the length of the BRTT was only 38% of the PSAT, this result confirmed theoretical expectations regarding the increased efficiency of adaptive over conventional testing.

A third major finding of the study was that p, the expected proportion correct over the entire item pool appeared to be the most desirable score for general use. Of all scores studied, θ, Ω, and p exhibited the highest reliabilities. However, forms A and B were not parallel with respect to θ, and Ω. However, when p was employed, the scores on both forms were directly comparable.

Chapter IV

Student Response to Adaptive Testing

4.1 Collection of Attitude Data

The process of performing an assessment involves two variables, the instrument and the individual. Chapter III was concerned with the properties of the instrument; this section is focused on the second variable--the student.

Siegel (1969) suggested that attitude judgments should be made immediately after the test session, because perceptions might be subject to motivated forgetting, which would reduce the initial differences in perceived validity. For the same reason, estimates of overall test difficulty and probability of success should also be made at that time.

Following completion of both forms of the BRTT, students were asked to complete a posttest questionnaire. One hundred twenty-four students completed the questionnaire (which is reproduced in Appendix A). The remaining students either pleaded fatigue or had insufficient time to complete the questionnaire within the allotted class period.

Attitudes are important to the extent that they affect performance. Weiner (1957) found that examinees with distrustful attitudes had impaired performance on the WAIS picture completion and similarities subtests. It was thought that the distrustful comments made by the examinee interfered with his ability to make the correct response.

I. Sarason (1972) and Wine (1971) have suggested that attitudes affect anxiety level and performance by distracting the examinee's attentional focus from task-relevant variables.

In the current study, the questionnaire employed included items designed to determine students' prior familiarity with computers (Koch and Patıence, 1977); subjective perception of difficulty, anxiety, and motivation (Prestwood, 1978); factors in the human/computer interface (Alderman, 1978); preference for adaptive vs conventional testing; and feedback or knowledge of ıesults (Prestwood, 1978). The questionnaire also solicited student opinion on the best and least liked factors in the adaptive test (Schmidt, Urry, and Gugel, 1977). These latter topics employed free-response incomplete sentence blank format; all other items were multiple-choice and were adapted from the studies refereıced above.

Since student attitudes are likely to vary as a function of the perceived importance of the test, and the test results did not affect the students' lives, the data reported must be viewed with caution. In particular, the levels of reported anxiety may be lower than that experienced in a "live" testing situation (Koch and Patience, 1977).

The section which follows summarizes previous work on student artitudes to computerized testing. Subsequent sections present the results for the specific variables measured by the questionnaire.

4.2 Previous Research on Student Attitudes to Computerlzed Tasting

There is a notable lack of literature investigating examinees' attitudes toward testing, computers, and computerized testing; but findings generally suggest favorable attitudes toward computerized adaptive testing. Ahl (1975) concluded from a Creative Computing Magazine survey tnat most Americans have a generally positive attitude toward computers and that two-thirds of the population has a fair understanding of the computer's role and function. Cartwright and Derevensky (1976) found that teacher education graduate students exposed to computer-assisted testing had more favorable attitudes toward computer-assisted instruction, toward programmed instruction, and toward the lectures than students not exposed to CAT. They suggested that positive experience with computerized tests can modify previous negative attitudes. Betz and Weiss (1976) found that for high-ability students, motivation was high on both stradaptive and conventional tests administered on a cathode ray terminal, while for low ability students it was high only on the stradaptive test.

Schmidt, Urry, and Gugel (1977) investigated attitudes of 163 individuals who took a computerized adaptive test of verbal ability and found an overwhelmingly positive response. Eighty-three percent of those responding preferred the adaptive test to pencil-and-paper testing; 69% felt the adaptive test was more fair than a conventional test. Only 10% of those responding indicated a preference for pencil-and-paper testing, and 4% felt that the adaptive test was less fair than a conventional test.

The features most liked about adaptive testing were: reduction in total test time (35%), simplicity of administration (19%), lack of time pressure (13%), and potential for quick feedback (10%). The features least liked were: the inability to review and change previous answers (23%) and difficulty in adjusting to this method of administration (20%) (Schmidt, Urry, and Gugel, 1977).

Hedl, O'Neil, and Hansen (1973) investigated preference for computer-administered vs examiner-administered intelligence tests. These investigators found that the computerized tests elicited higher anxiety and less favorable attitudes than the examiner-administered test. However, these results were probably due to tne fact that the computerized protocol (which was nonadaptive) required that all examinees complete the entire test, whereas the examiner-administered test was terminated after the individual failed 10 consecutive items. Given the massive failure experience, the resultant negative attitudes are not surprising. It is interesting to note that failure feedback on an "intelligence test" is a standard procedure for experimentally inducing test anxiety (Levitt, 1967).

Koch and Patience (1978) investigated student attitudes toward tailored achievement testing. The variables they measured were (1) time pressure, (2) perceived test difficulty, (3) test anxiety, (4) prior experience with computers, and (5) overall preference for computerized vs conventional testing.

These investigators compared attitudes under two circumstances: (1) a condition in which the test did not count toward the course grade, and (2) a condition in which the test did count. Results indicated that students felt significantly more time pressure and anxiety under circumstances in which the test counted toward the final grade, but no significant differences were found with regard to perceived test difficulty or overall preference for adaptive vs conventional testing.

A major advantage of computerized test administration is the ability to provide the examinee with feedback or knowledge of results (KR). Research on KR in an adaptive testing environment has been conducted by Betz and Weiss (1976a, b), Pine (1977), and Prestwood (1977). These findings will be discussed in section 4.5, which is concerned with feedback. It appears that KR affects a number of attitudinal variables. Betz and Weiss (1976a, b) found that accuracy of perception of test difficulty and motivation were higher for students in a KR condition than for students who received no feedback.

4.3 Prior Familiarity with Computers

Some individuals in our society have expressed distrust and unease with computer technology. To the extent that such attitudes adversely affect test performance they are of interest in adaptive testing. Koch and Patience (1977) have suggested that items tapping prior familiarity with computers can serve as useful covariates in analysis of subsequent questionnaire data.

The current study employed three items adapted from Koch and Patience's (1977) study. These items asked if the student was at all familiar with computers, had used a keypunch or terminal before. The responses to these items are presented in Table 4.3a. As can be seen from the table, about half the students reported some familiarity with computers and the same number had interacted with computers by means of a terminal. Students were more familiar with computer terminals than with keypunch machines, a finding that may be surprising to those whose familiarity with computer systems was acquired during the time when "batch" rather than interactive systems were predominant.

4.4 Perception of Test Difficulty, Anxiety, and Motivation

The variables considered in this section are important determinants of an individual's performance. The importance of anxiety as a factor in task performance has been of major concern since Mandler and Sarason's (1952) seminal article on test anxiety. A review of the test anxiety literature is beyond the scope of the present work; however, anxiety has

Table 4.3a

Prior Familiarity with Computers. Total Sample (N=124).

Item	Text of Item	Response to Options	
		N	%
14.	Are you at all familiar with computers?		
	Yes	60	48%
	No	63	51%
	Omit	1	1%
15.	Have you ever punched computer cards at a keypunch machine before?		
	Yes	44	35%
	No	80	65%
16.	Have you ever interacted with a computer by means of a terminal before?		
	Yes	65	52%
	No	59	48%

been repeatedly shown to adversely affect performance at all levels of academic experience (Gaudry and Spielberger, 1971). Many theorists have noted that, since anxiety can serve as a learned drive, it can serve as a motivational variable and will interact with other motivational variables. One of the most relevant of the motivational theories is Atkinson's (1958) model of fear of failure, need for achievement, and their interactional effect on risk-taking behavior.

Most risk-taking experiments investigate risk-taking as a voluntary action (a dependent variable). In general, individuals with high motive to approach success prefer to answer questions at an intermediate level of difficulty, while individuals with high motive to avoid failure prefer to answer questions that are either very easy or very difficult (Atkinson, 1958). In an adaptive test, however, the level of uncertainty is imposed by the selection algorithm. Disregarding guessing, all the questions are aimed at the .5 probability of success for a given examinee; these are precisely the type of questions that the examinee with high fear of failure would tend to avoid. High motive to avoid failure has been equated with high test anxiety (Atkinson and Litwin, 1960). Consequently, an adaptive test might be stressful for a high-anxious testee who would be predicted to experience maximum avoidance tendencies on items which have a .5 probability of success. Anxiety caused by these avoidance tendencies might be hypothesized to adversely affect performance. In particular, it could encourage impulsive responding in order to simply remove the anxiety-producing stimulus; or it may lead to reflective responses when the examinee is unable to choose among competing responses (Spence, 1964). However, Atkinson (1958) predicted that when there is no choice of level of task difficulty, performance should be optimal when the probability of success is .5 regardless of whether the motive to a hieve or the motive to avoid failure is stronger. This is hypothesized to occur because both the motive to achieve and the motive to avoid failure would be greatest at the 50% point, and both effects would summate, resulting in maximum motivation to perform. While this theory predicts maximum performance on the basis of maximum motivation to perform, it is possible that anxiety caused by such a situation would be sufficient to interfere with the positive relationship between motivation and performance (Wine, 1972).

Atkinson's theory suggests that adaptive testing might be more stressful to individuals with high test anxiety than conventional tests. However, a conventional test which is too difficult for an individual is likely to result in a high proportion of guessing. Guessing, because of its high random component, reduces the measurement accuracy of a test. The problem is even more complex when a student guesses as a result of partial knowledge which permits one or more distractors to be eliminated (Lord and Novick, p. 303).

While only a small portion of testees generally find a test to be entirely too difficult, a considerable portion of testees experience short groupings of overly difficult questions. After a failure experience with difficult problems, the examinee may develop an impulsive response set which may lead to errors on subsequent items of more appropriate items. Walker, Neilson, and Nicolay (1965) found that under stress conditions (caused by failure at a previous task), anxiety was negatively correlated with intelligence test performance. Hedl, O'Neil, and Hansen (1973) found that subjects had greater anxiety and more negative attitudes toward computer-based testing after a massive failure experience with items that were too difficult.

Clearly the issues of the examinee's anxiety, motivation, and perception of test difficulty are important and their effect on an individual's performance are complex. Unfortunately, it is difficult to obtain data on these issues during an important testing session, since the introduction of experimental interventions during testing is rarely possible. Since the data collected in the present study were obtained from student volunteers in an experimental context, they must be regarded with caution. However, because of the importance of these variables it was decided to collect relevant data. To facilitate comparison of these data with related research, items were adapted from previous studies.

Items 17-22 on the questionnaire were concerned with the student's perception of the difficulty of the adaptive test. These items were adapted from a study by Pine (1977, personal communication) and are reported in Table 4.4a.

Items 17 and 18 were concerned with the students' perception of the appropriateness of the difficulty of the test. Inspection of the response distributions reveals that almost none of the students found the test items always or frequently too easy. Ninety percent of the students found them sometimes or seldom too easy. Eighty-five percent of the students found the items too difficult sometimes or frequently; however, only 19% of the students indicated that they guessed more than half the time. The picture that emerges from these responses is that the students generally felt that the test was appropriate to their ability level, although somewhat on the difficult side. Very few students found the test exceptionally easy or difficult.

Item 21 asked the students to rate the difficulty of the test, overall, in relation to their ability. Half the students felt that the test was "just about right" while most of those remaining found it "somewhat too difficult."

It is interesting to speculate to what extent the students' expectations of success influenced their judgment of difficulty. Students of high ability generally achieve high number-right scores on conventional tests while those of low ability generally obtain low number-right scores. In the adaptive test, the number-right score should be unrelated to ability level. In any case students generally expressed low levels of frustration with the test.

Table 4.4b presents the response distribution for items related to the students' self-reports of anxiety during testing. Generally, students reported moderate levels of worry; but few (3%) felt that anxiety unquestionably prevented them from doing their best, while 12% felt that anxiety might have affected their scores somewhat. Prestwood (1977) found that examinees who took tailored tests which counted toward a course grade reported higher levels of anxiety than those whose tests did not count. It is probable, therefore, that the results of the current study are not indicative of the anxiety levels which would occur if the test were important in the individual's academic career.

Table 4.4c presents the response distributions for a series of items adapted from Pine (1977), designed to measure motivation. As can be seen from the responses to items 28 and 29, about 30% of the

Table 4.4a

Subjective Perception of Difficulty. Total Sample (N=124).

Item	Text of Item	Response to Options N	%
17.	How often did you feel that the questions in the test were too easy for you?		
	Always	0	0%
	Frequently	6	5%
	Sometimes	65	52%
	Seldom	47	38%
	Never	5	4%
	Omit	1	1%
18.	How often did you feel that the questions in the test were too hard for you?		
	Always	1	1%
	Frequently	35	28%
	Sometimes	71	57%
	Seldom	17	14%
	Never	0	0%
19.	On how many of the questions did you guess?		
	Almost all of the questions	1	1%
	More than half of the questions	4	3%
	About half the questions	18	15%
	Less than half of the questions	55	44%
	Almost none of the questions	46	37%
	None of the questions	0	0%

Table 4.4a (Continued).

Subjective Perception of Difficulty. Total Sample (N=124).

Item	Text of Item	Response to Options N	%
20.	How often were you sure that your answers to the questions were correct?		
	Almost always	2	2%
	More than half of the time	31	25%
	About half of the time	54	44%
	Less than half of the time	33	27%
	Almost never	3	2%
	Omit	1	1%
21.	In relation to your vocabulary ability, how difficult was the test for you?		
	Much too difficult	1	1%
	Somewhat too difficult	53	43%
	Just about right	65	52%
	Somewhat too easy	4	3%
	Much too easy	0	0%
	Omit	1	1%
22.	Did you feel frustrated by the difficulty of the test questions?		
	Not at all	39	31%
	Somewhat	79	64%
	Fairly much so	6	5%
	Very much so	0	0%

Table 4.4b

Anxiety. Total Sample (N=124).

Item	Text of Item	Response to Options	
		N	%
23.	During testing, did you worry about how well you would do?		
	Not at all	33	27%
	Somewhat	64	52%
	Fairly much so	20	16%
	Very much so	7	6%
24.	Were you nervous while taking the test?		
	Not at all	81	65%
	Somewhat	31	25%
	Moderately so	12	10%
	Very much so	0	0%
25.	How did you feel while taking the test?		
	Very tense	1	1%
	Somewhat tense	14	11%
	Neither tense nor relaxed	43	35%
	Somewhat relaxed	43	35%
	Very relaxed	23	19%
26.	Did nervousness while taking the test prevent you from doing your best?		
	Yes, definitely	4	3%
	Yes, somewhat	15	12%
	Probably not	71	57%
	Definitely not	34	27%

Table 4.4c

Motivation. Total Sample (N=124).

Item	Text of Item	Response to Options	
		N	%
27.	How frequently were you careful to select what you thought was the best answer to each queation?		
	Almoat always	66	53%
	Frequently	36	29%
	Sometimes	20	16%
	Rarely	2	2%
	Never	0	0%
28.	Did you feel challenged to do as well as you could on the test?		
	Not at all	8	6%
	Somewhat	41	33%
	Fairly much so	40	32%
	Very much so	35	28%
29.	Did you care how well you did on the test?		
	I cared a lot	39	31%
	I cared some	62	50%
	I cared a little	14	11%
	I cared very little	7	6%
	I didn't care at all	1	1%
	Omit	1	1%

students expressed high levels of motivation. Half of those taking the test indicated that they were careful to choose the correct answer almost every time. As with anxiety, it would be interesting to compare these results with data obtained following a testing session of personal significance to the student.

4.5 The Role of Feedback (Knowledge of kesults)

It is relatively simple to provide immediate feedback in an adaptive testing system. However, because the effects of feedback on performance appear to depend on complex interactions, it is not clear under what circumstances feedback would facilitate or impair the performance of the examinee.

Betz and Weiss (1976a, b) studied motivation, anxiety, and performance as a function of provision of knowledge of results (KR) for high- and low-ability examinees on adaptive and conventional tests. High-ability examinees, overall, reported more motivation than low-ability examinees. KR resulted in increased motivation for high-ability examinees and decreased motivation for low-ability examinees. Furthermore, motivation was higher on the conventional test for high-ability examinees (where KR was probably mostly positive), and higher on an adaptive test for low-ability examinees (where KR was probably more positive than for a traditional test). In contrast, Means and Means (1971) found that high-ability students performed better with negative KR; and low-ability students performed better with positive KR. However, in this case KR was given after the entire test. Item by item, KR is psychologically quite different from posttest KR. For example, a student who receives negative feedback on the first few items of a test may give up at the beginning. If the KR is provided after the test, the examinee may be motivated to achieve a higher level of performance on a subsequent test. This distinction is supported by Locke et al. (1968), who found that the motivational effects of KR depend on the goals the examinee sets in response to the KR.

Betz and Weiss (1976a, b) also found that high-ability examinees reported less anxiety than low-ability examinees on the same type of test. KR produced higher anxiety for the adaptive test and lower anxiety for the conventional test for both ability groups. Hansen (1974) found that high-anxious testees made more errors with feedback than without it. Weiner and Adams (1974) found evidence that failure and the anxiety it can induce may lead to more reflective responding on a matching familiar figures test. Hansen (1974) also found that while feedback helped the performance of high-reasoning testees, it impaired the performance of low-reasoning testees.

The most striking finding of the Betz and Weiss (1976a) study was that KR led to significant increases in test scores for the total group of examinees. KR yielded greater performance improvement on the conventional, as compared to the adaptive, test.

Prestwood (1977) studied the effects of KR on 561 undergraduates using a modified stradaptive algorithm which yielded tests of high (40% correct), medium- (60% correct), or low-difficulty (80% correct). Three conventional peaked tests were constructed to yield comparable mean number-right scores. This study failed to replicate Betz and Weiss's (1976a) finding of better performance in the KR condition.

Table 4.5a

Feedback. Total Sample (N=124).

Item	Text of Item	Response to Options	
		N	%
34.	Would getting feedback on the test make it:		
	More interesting	95	77%
	Less interesting	4	3%
	Cannot say	25	20%
35.	Would getting feedback after each question make you nervous?		
	Very nervous	17	14%
	Somewhat nervous	29	23%
	Slightly nervous	45	36%
	Not nervous at all	32	26%
	Omit	1	1%
36.	How would you feel about getting feedback?		
	I would rather not know whether my answers were right or wrong.	14	11%
	I really don't care if I got feedback or not.	5	4%
	I would like getting feedback after each question.	33	27%
	I would like feedback at the end of the test.	72	58%

Nor did Prestwood find higher anxiety on the adaptive test as reported by Betz and Weiss.

It appears as though the effects of KR on performance and attitude are extremely complex. Strang and Rust (1973) pointed out that in order to test the effect of KR, it is necessary to control for intrinsic KR of ongoing activity. If the examinee can estimate performance, KR will be redundant. Controlling for intrinsic KR, Strang and Rust found that the examinees were more nervous with immediate KR than without it. Betz and Weiss (1976b) found that even though examinees taking the adaptive test were more nervous with KR, 90% of all examinees in the study said they liked the provision of KR. The actual proportion of positive KR was related to attitude toward the test: the greater the proportion of positive KR, the more favorable the examinee's attitude toward the test. Prestwood (1977) also found that a high proportion of examinees liked having KR.

In the current study, KR was not provided, although the design of the CAT system makes provision of both item and total test feedback a simple matter. The reason for not employing feedback was that the goals of the study were to explore the performance of the BRTT, and feedback would have added a confounding variable.

However, it was decided to include three items to which students could respond by indicating their preferences regarding feedback. These items were adapted from Prestwood's (1977) scale and are presented in Table 4.5a. Three-quarters of the students felt that getting feedback would make it more interesting. Ninety-nine percent of the students in Prestwood's study felt that feedback made the test more interesting.

Although 76% of the students in Prestwood's study felt that feedback after each item did not make them nervous, only 26% of the subjects in the present study agreed; 14% felt that item feedback would make them very nervous.

4.6 Human Factors in the Computer/Human Interface

The ease with which the student is able to interact with the computer is a critical factor in the testing experience. In the design of the computer system employed in the present study, a careful effort was made to develop a simple interaction protocol which would be intuitive in operation and low in fatigue. Two factors in the design were the development of a customized keypad to eliminate the need to "hunt and peck" on a typewriter keyboard and the design of terse instruction sets which were employed after the student had been exposed to the complete, verbose instructions for a given item type. The use of terse instructions reduced the reading load required.

Table 4.6a presents the results of two items (adapted from Alderman, 1978) dealing with the human/computer interface and fatigue.

4.7 Preference for Adaptive vs Conventional Testing

Table 4.7a indicates that a majority of students who had a preference would prefer adaptive to conventional testing. It is difficult to determine how seriously this preference would transfer to an actual testing environment; the novelty and experimental nature of the study are doubtless biasing factors. However, the lack of a strong negative response to this question and others (such as difficulty, anxiety, and motivation) suggests that students will be receptive to adaptive tests

Table 4.6a

Human Factors. Total Sample (N=124).

Item	Text of Item	Response to Options	
		N	%
30.	Did the mechanics of using the computer terminal interfere with your taking the test:		
	Not at all	78	63%
	Slightly	34	27%
	Somewhat	10	8%
	Very much so	2	2%
31.	How tiring did you find the computer-administered test?		
	Very tiring	3	2%
	Somewhat tiring	15	12%
	Slightly tiring	46	37%
	Not tiring at all	60	48%

as compared to conventional tests. The free responses to the incomplete sentence blanks which follow may help clarify factors which affected the students' response to the adaptive testing situation.

Table 4.7a

Preference for Adaptive vs Conventional Testing
Total Sample (N=124).

Item	Text of Item	Response to Options	
		N	%
32.	Compared to a "paper-and-pencil" multiple-choice test of the same length would you:		
	Find the computer test more tiring?	13	10%
	Find both tests about the same?	41	33%
	Find the paper-and-pencil test more tiring?	69	56%
	Omit	1	1%
33.	If you had a choice, would you prefer to take the PSAT as:		
	A computer-administered test	58	47%
	A pencil-and-paper test	39	31%
	No preference	26	21%
	Omit	1	1%

Table 4.7b

Responses to Open-ended Items

I. What did you like best about the computer-administered test?

1. More interesting to take.

2. Less intimidating.

3. Less time consuming.

4. Not as tiring as paper-and-pencil test.

5. I prefer not looking ahead to other questions.

6. Easy to use and understand.

7. Less chance of making errors.

8. Taking the test alone (or with 1 or 2 others) is more private, so you feel less pressured.

9. I liked not being rushed, and being able to take my time.

10. I liked not being interrupted by a Proctor.

11. The directions were clear.

12. There were no essay or math queations.

II. The thing I like least about this method of administering an examination is:

20. Not being able to go back to the previous question/answer to either review or change.

21. Some of the lettera were difficult to read, making concentration harder.

22. Not knowing your score on the teat.

23. The delayed time between the anawer and the next question.

24. It waa distracting for each question to be printed out letter by letter.

25. The computer took aome time to get ready.

26. The directiona could have been a lot shorter.

Table 4.7b (continued)

Responses to Open-ended Items

27. The test was time-consuming and boring.

28. The screen bothered your eyes at times (possibly causing headaches).

29. Just staring at the screen became annoying.

30. The questions seemed harder to me than a written test.

31. Haphazard guessing was caused due to the faster, more relaxed test.

32. Felt I was being rushed.

33. Became reckless by the end of the test.

34. Only using 3 keys on the keyboard.

35. Having to press the "enter" key twice when you had no changes.

36. Takes time adjusting to use the computer.

37. Computers make me more nervous than the pencil-and-paper test.

38. Computers are being too widely used.

39. Not private enough, someone could read your answers if desired.

40. The location of the computers.

41. The uncomfortable chair.

42. Worried if something could go wrong with the computer.

43. I found the test too easy.

44. I found nothing wrong with the test, I really liked it.

III. Changes to consider about this method of administering and examination:

50. The inability of the user to review his answers and to make changes if necessary.

51. Being able to tell the students what the results of the test are.

52. To see the whole test together at the end of the test.

Table 4.7b (continued)

Responses to Open-ended Items

53. The computers response should be quicker.

54. The repeating of instructions.

55. After each question completed, there should be a print-out to how many questions have been completed and how many left.

56. The visibility of the print-out should be more legible.

57. The letter "G".

58. Use a diffcrent letter style for the computer's print-out.

59. The whole question should appear at once on the screen, not printing-out letter by letter.

60. Getting practice using the computer before taking the test.

61. Don't believe in "Practive-test", the test should count.

62. The test should be taken in total privacy.

63. Questions dealing with different subjects should be considered.

64. Use the same kind of questions in a consecutive order.

65. The buzzing noise.

66. The chair!

67. The seating arrangement.

68. The place.

69. The machine should make some noises.

70. A way to make the student feel more relaxed.

71. Nothing has to be changed.

Chapter V

Implications and Recommendations

5.1 Overview

The current study has demonstrated the viability of adaptive testing in the high school environment. From an operational point of view, the system performed reliably and students found interaction with the computer terminai to be a simple task. Psychometrically, the performance of Broad-Range Tailored Test was generally consistent with theoretical expectations. A summary of the major findings of the study will be found in section 3.13. The present chapter considers the implications of the study for future development efforts.

The chapter is organized around four recommendations. The recommendations are:

1. That the organization collaborate with an interested client to develop an adaptive test for use in an educational setting.
2. That the potential for microprocessor-based systems for the delivery of adaptive testing be evaluated.
3. That extensions to item response theory and the development of alternative models for the provision of adaptive testing be explored.
4. That high priority be accorded the development of innovative assessment strategies for computer presentation. Such items might involve simulation and gaming, constructed responses, graphics, motion, sound, and time-dependent responses.

The four recommendations involve both development and research components. Underlying these four recommendations is the belief that adaptive testing has reached a state of development in which its practical applications can be seriously contemplated. Although many areas exist in which additional research is needed, research agendas will benefit from the experiences of developing an operational system. A second belief which underlies the recommendations is that the development of operational adaptive testing will be facilitated by collaboration among educators, technical staff, test development specialists, and psychometricians.

Sections 5.2 through 5.5 elaborate upon the four recommendations enumerated above.

5.2 Recommendation 1: That the organization collaborate with an interested client to develop an adaptive test for use in an educational sett ng.

Recommendation 1 proposes that the organization continue its work in adaptive testing with the development of a test for use in an educational setting. Although all of the research problems related to adaptive testing have not been solved, there is much to be learned from a modest operational project. Since the present study as well as previous studies have supported the major theoretical predictions of adaptive testing models, the development of a valid and reliable opera ional instrument is a reasonable goal.

In selecting an operational project, several factors will be important. A creative, flexible project team and a well-defined set of goals will be crucial to the success of the project. The project should be one in which the adaptive test fills a need which cannot readily be met by conventional paper-and-pencil testing. The project should be modest in its goals and be designed to facilitate future development efforts through the collection of data appropriate to both formative and summative evaluation.

A number of areas appear to hold promise for adaptive testing. In the community college environment, for example, adaptive testing could be used in conjunction with walk-in registration. Students desiring to take English or mathematics courses could respond to a relatively short test administered on a computer terminal which would determine the appropriate placement for the student. Placement testing could be integrated with the registration procedures thus providing an effective means of tracking large numbers of part-time students.

A second area in which adaptive testing appears to hold promise is that of diagnosis. In order to be effective, diagnostic batteries must be comprehensive in scope. When a large number of characteristics are to be tested the number of items which must be administered can become unreasonably large. Because of its efficency, adaptive testing would be an effective alternative to the administration of batteries of conventional tests. Multistage procedures in which routing tests are used to perform gross discriminations and branch individuals to more molecular assessment segments are feasible. Diagnostic testing might be used with special populations such as the disadvantaged or in special education settings. In conjunction with remedial programs such diagnostic testing could be integrated with instructional modules to create a comprehensive, automated mastery learning environment.

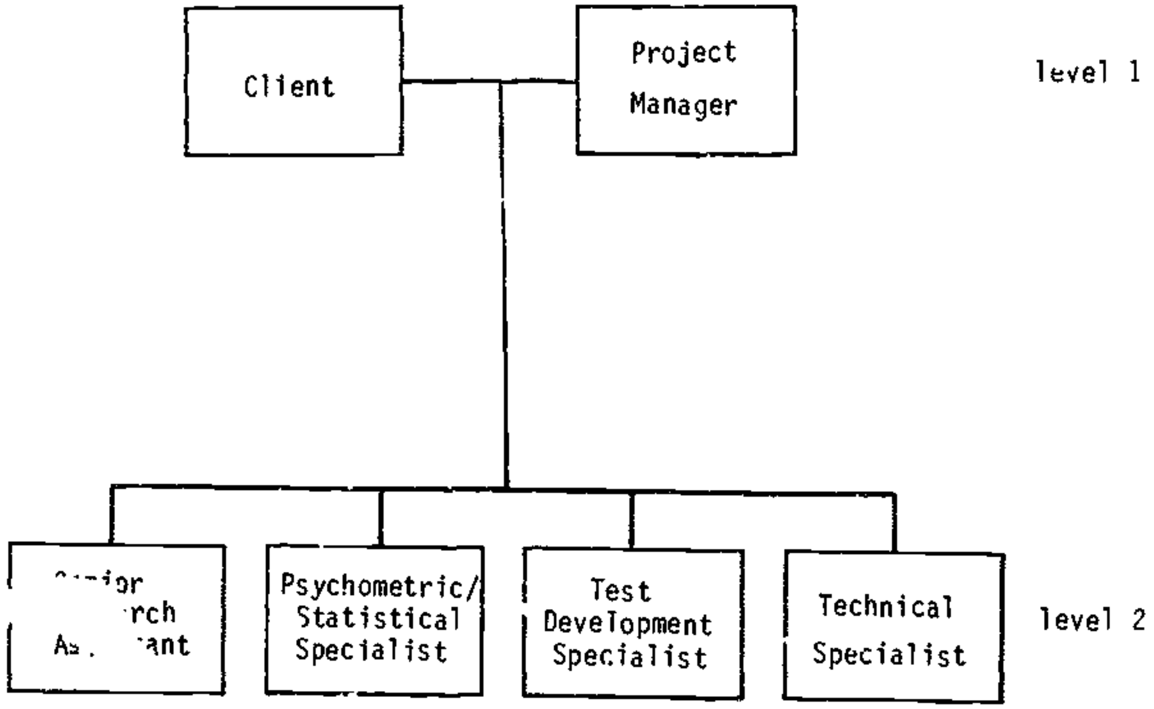

Client
Project Manager
level 1
Senior Research Assistant
Psychometric/ Statistical Specialist
Test Development Specialist
Technical Specialist
level 2

Figure 5.2a Staff for an Operational Development Project

At the present time the use of adaptive testing for selection should be approached with caution. The advent of truth-in-testing legislation may place constraints upon the security of the item pools used in adaptive testing. This is conceptually no different from the problems facing other testing programs. One solution might be to employ very large item banks which could be published; the number of items in the pool would have to be sufficiently large to prevent an individual from memorizing the responses to the entire pool. However, this technique would require considerable quantities of direct-access storage and may be impossible to implement on current microprocessor-based systems.

The program manager must have an understanding of measurement, computer technology, and educational practice. One important function the program manager would serve is to facilitate the conceptualization of the project in terms which are meaningful to both the client and the project staff. As coordinator, communicator, and facilitator, the program manager would maintain the project's momentum and direction.

The program manager must have sufficient technical expertise to manage the technical aspects of the project. It is important that s/he be able to evaluate technical alternatives and be able to communicate with both technical and nontechnical individuals. A lack of communication between technical and subject-matter experts is a common cause of frustration and failure in projects of this nature.

The four specialists working on Level 2 will share responsibility for design and implementation of the system. The technical specialist would be responsible for system design, programming, and hardware selection. Because technical subtleties are often mysterious to nontechnical individuals, the project team will place heavy reliance upon the technical specialist for support; this is especially important if the technical leader and technical specialist roles are combined.

The test development specialist would be responsibile for the test specifications and items tor the adaptive test. This individual should be knowledgeable in item characteristic curve theory. A test development specialist who is reasonably familiar with the capabilities of computers would tend tc be more creative than one to whom computers are unfamiliar.

The role of the statistical/psychometric staff member is a crucial one. He or she would be responsible for designing the mathematical foundation on which the test is constructed. These tasks include: determination of the selection-algorithm, development of the item pool structure, calibration of items, choice of stopping rule, selection of numerical analytic techniques, determination of score transformations, developmen of equating methodologies for alternate forms, and the conduct of simulation studies to validate the performance of model. He or she should also be well versed in item characteristic curve theory and should have a good knowledge of previous research in computerized adaptive testing. It would be helpful if he or she had some programming background, particularly in the area of numerical analysis.

The Senior Research Assistant would be responsible for maintaining item files, helping prepare the system documentation, assisting in the preparation of user manuals, and generally providing support for the wide range of administrative functions which a project of this nature requires.

5.3 Recommendation 2: That the potential for microprocessor-based systems for the delivery of adaptive testing be evaluated. This evaluation should include three models: a time-sharing model; a stand-alone microprocessor model; and a network model in which a host mainframe computer supports a network of microprocessors.

Over the last decade there has been a tremendous increase in the sophistication of design and fabrication techniques for electronic circuitry. The class of circuits which have resulted from these new technologies, known generically as microelectronic components, have been used as a variety of applications ranging from digital watches and hand-held calculators to electronic computers and satellites. Microelectronic circuits are characterized by a high degree of integration. Thousands of transistors and other circuit components are fabricated on a thin silicon wafer or "chip" whose measurements are typically .16" x .22" (Noyce, 1977). It is difficult to overestimate the impact of microelectronic developments on computer technology. The microelectronics revolution is far from over and the technology is advancing at an unprecedented rate.

To appreciate the magnitude of the size and cost reductions which have occurred as a result of microelectronics, compare ENIAC (Electronic Numerical Integrator and Calculator)--the first electronic computer with a typical microprocessor. Designed by Eckert and Mauchly at the University of Pennsylvania and operational in 1946, ENIAC weighed 30 tons, required 150 kilowatts of electricity, and contained in excess of 20,000 vacuum tubes. ENIAC was capable of multiplying two 10 digit numbers in 0.003

seconds; its memory size was 150 locations. Reliability was a major concern. With over 20,000 vacuum tubes in ENIAC, the rate of random tube failure approached the time required to locate and replace the malfunctioning tube. This posed a serious problem to future development since it was hypothesized that computers using larger numbers of vacuum tubes would rapidly approach zero operational time due to the large number of failures expected.

In contrast, consider a typical microprocessor chip, INTEL Corporation's 8085 Microprocessor. The 8085 is fabricated on a single chip which measures .164" x .222". The chip contains 6,200 transistors and is capable of decoding over 300 instructions. It can execute 770,000 instructions per second. The manufacturing cost of an 8085 chip is measured in pennies; its retail cost is several dollars. The chip is rugged, reliable, and may be powered by batteries.

It is evident that the computer is no longer an expensive laboratory device and the availability of microprocessors profoundly alters the cost/benefits of applying educational technology to the classroom. Cost reduction trends are expected to continue as manufacturing technology becomes increasingly sophisticated. As Figure 5.3a shows, the number of components per circuit has doubled every year since 1959. Figure 5.3b shows the actual and projected decline of the cost/bit of computer memory for the years 1973-83.

At the time that the CAT system was designed, general purpose microcomputers were not readily available. For this reason, the CAT system was designed to run upon a conventional time-sharing system. Figure 5.3c illustrates the structure of the time-sharing system employed for the CAT project. As can be seen from this illustration, a single central computer services multiple students using a single set of items maintained on magnetic disks. Communication between the student at the terminal and the central computer occurs across telephone lines.

At the present time, general purpose microprocessor systems have become available, among them some designed especially for educational use. Unlike time-shared systems, each user of a microprocessor system has sole use of the processor. In a microprocessor system, the actual computer circuitry represents a small fraction of the total cost; the most expensive components tend to be such items as the keyboard, disk drive, and display tubes. For this reason it is rarely economical to time-share microprocessor systems. Figure 5.3d illustrates a typical microprocessor configuration which might be employed for adaptive testing.

Microprocessor systems have both advantages and disadvantages as compared to conventional time-shared systems. The major disadvantages of microprocessor systems are that they do not yet have the storage capabilities typical of large-scale processors, and their computational power, although impressive, tends to be less than that of large scale systems. In effect, microprocessors are scaled down, compared to larger processors. The disadvantages of smaller-scale processors, however, are balanced by a number of important advantages. A major advantage of microprocessor systems is that they cost far less than large-scale systems. It is possible to purchase a microprocessor system with 48K bytes of memory and a "floppy" disk for about $3,000. In contrast, the retail price of the PDP-11/40 computer used in the current study was approximately $150,000. It is evident that a large number of microprocessors can be purchased for the cost of a single large-scale system.

A second advantage of microprocessor systems is that hardware malfunction affects only a single testing station. In a time-shared system, failure of the central processor causes all terminals to stop operating. A third advantage of the microprocessor-based system is that its operating costs tend to be lower. Large-scale systems generally require the attention of an operator at the central site, whereas in microprocessor-based systems the user serves as operator. Additionally, since the microprocessor is located at the testing site, telephone lines are not needed for data transmission--a factor which can result in considerable cost savings.

As shown in Figure 5.3d, the microprocessor can support low-cost input/output devices which may facilitate testing. For example, the microprocessor can control a tape recorder or speech synthesizer to provide audio stimuli. A light pen would permit the testee to point to the chosen option or to part of a diagram. Image storage devices such as slide projectors or microfiche can provide randomly accessed graphics.

Although microprocessors appear to offer considerable advantages for adaptive testing, it is not feasible simply to transfer the current CAT system to a microprocessor. One reason for this is that the storage capabilities of "floppy" disks are considerably less than those of the disks currently used. Analysis needs to be performed in order to determine whether the storage capabilities of microprocessor systems are adequate to meet the needs of an adaptive testing environment. In addition, analysis of the numerical algorithms employed in the current systems to determine their transferability to microprocessor systems needs to be undertaken. It is probable that current microprocessor systems will prove capable of supporting adaptive testing. Urry (1979) has demonstrated an adaptive verbal ability test which employs a microprocessor.

Figure 5.3e illustrates the design of a system which employs a network consisting of a central host processor and remote microprocessor testing stations. In this system the microprocessor stations function as independent testing stations as in the microprocessor model (Figure 5.3d) but also have the capability of two-way communication with the central processor. Using this communications capability the microprocessor could transmit registration and item response data to the central computer for score reporting and item analysis. For example, a student might take an adaptive test which includes several experimental items. The item responses would be transmitted to the central computer. Test items could be achieved for subsequent score reporting and responses to the experimental items could be used for item analysis. Since the network could also support communication from the central processor to the microprocessor station, the central facility could transmit new tests to the microprocessor.

It is evident that the microprocessor has considerable potential as a vehicle for adaptive testing. It is therefore recommended that the organization evaluate the potential of microprocessor-based systems for the delivery of adaptive testing. As part of these evaluations it is suggeste' that two hardware technologies of high potential significance--bubble memory and image processing--be examined.

Bubble memory exploits a recently discovered property of certain crystaline materials. These materials have the characteristic that, when a wafer is magnetized in a direction perpendicular to the plane of the largest surface, magnetic zones known as "bubbles" appear. Bubble memories can do everything that disks do; that is, they can store large amounts of data and provide random access to any portion of the data. But unlike rotating magnetic disks, bubble memory performs these functions electronically rather than mechanically. Unlike disks, bubble memories have no motors, rotsting magnetic surfaces, or moving heads and would offer significant advantages in an educational environment in which rough handling and a lack of trained maintenance personnel are the norm. Further, unlike conventional computer memories, bubble memory is nonvolatile; the data in it are maintained even when power to the system is turned off. It is predicted that bubble devices will store data extremely accurately, and without loss, for over a century. Because of its relatively low cost and high reliability, bubble momory appears to be a potential storage medium for item banks. Some bubble memory chips are now commercially available and have been employed in text processing systems.

A second technology of major importance to adaptive testing is image processing. Traditionally, computers have been extremely useful for processing text material but have not had the capabilities of storing and

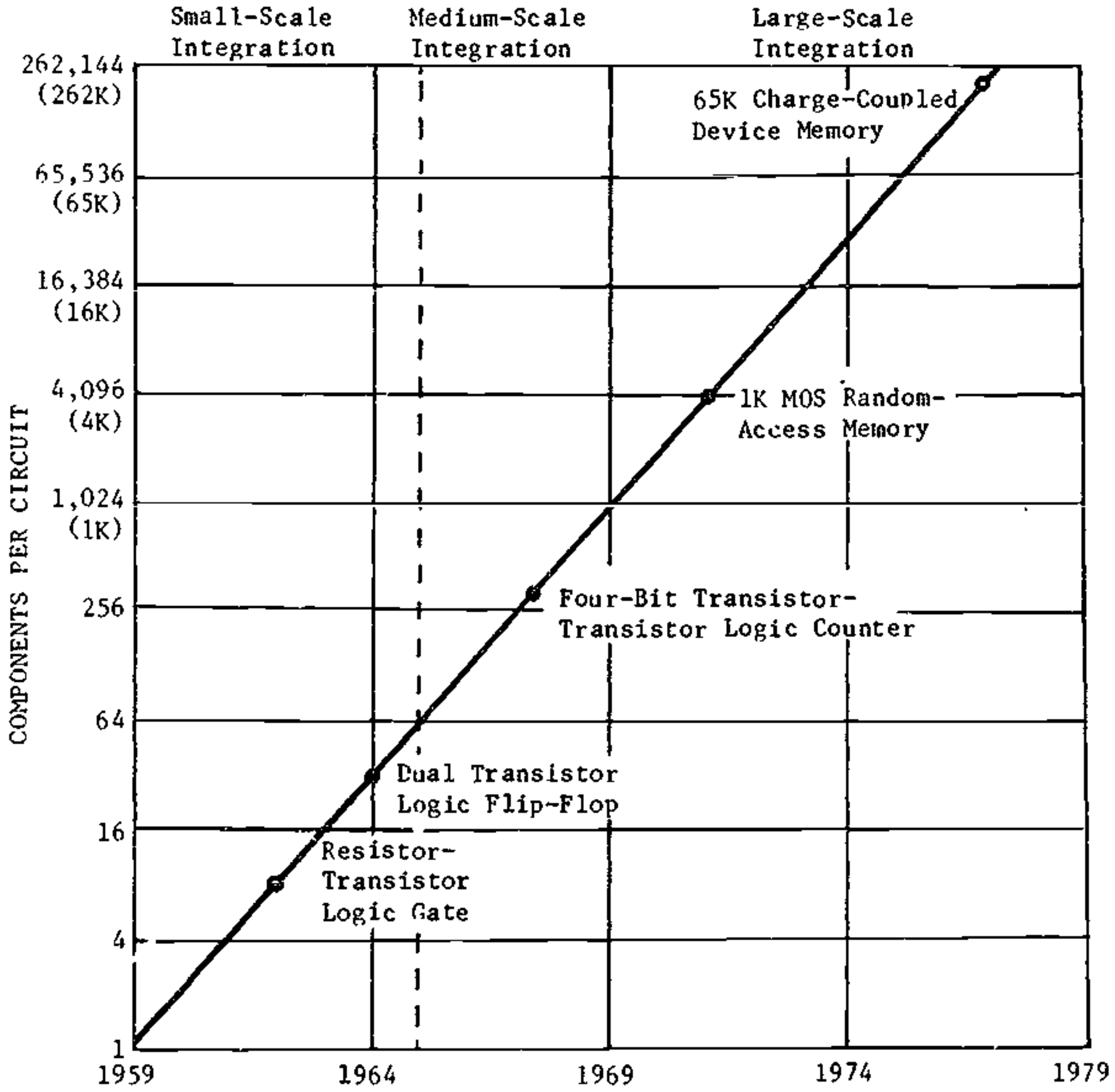


Figure 5.3a Number of Components/Circuit 1959-1979

From Microelectronics by Robert N. Noyce. Copyright © 1977 by Scientific American, Inc. All rights reserved.

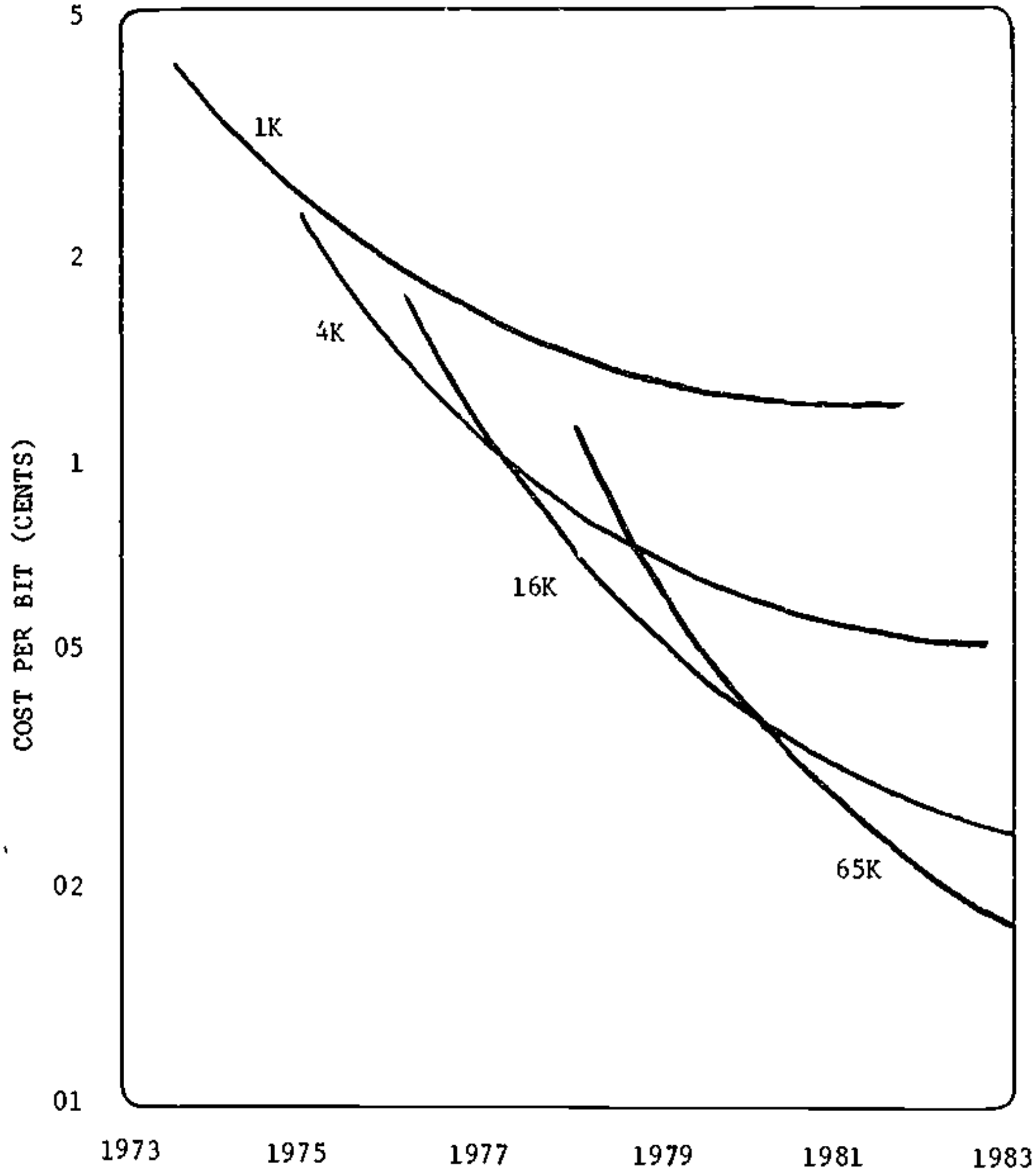


Figure 5.3b Actual and Projected Cost/Bit of Computer Memory 1973-1983

From Microelectronics by Robert N. Noyce. Copyright ©1977 by Scientific American, Inc. All rights reserved.

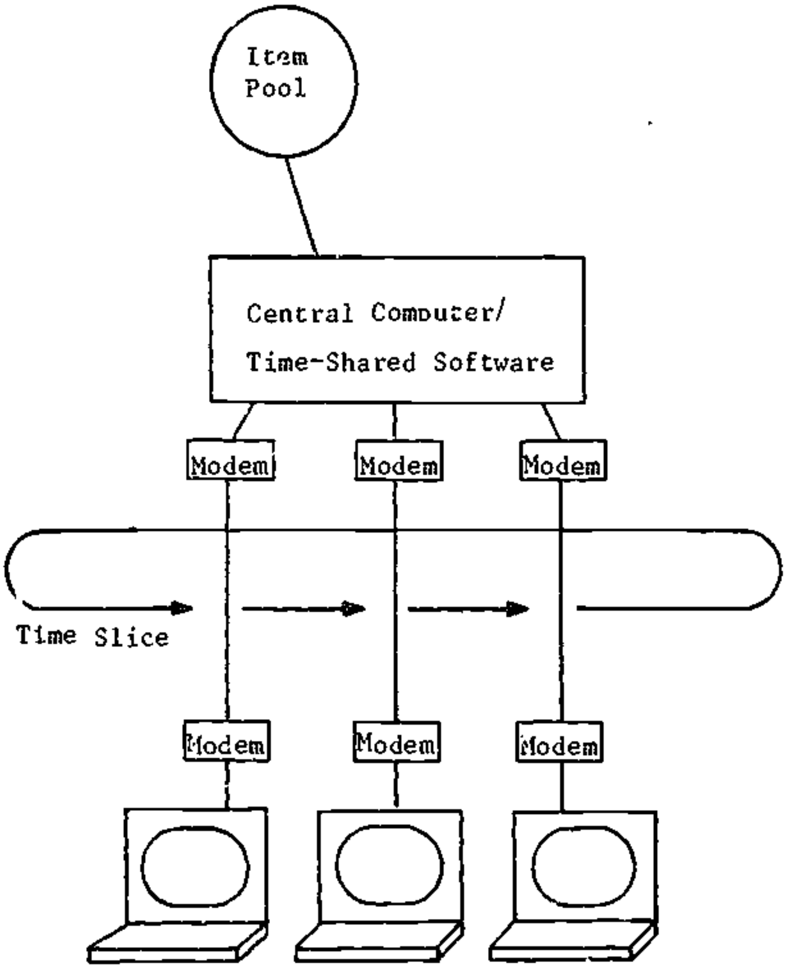


Figure 5.3c The Time-Shared CAT Model

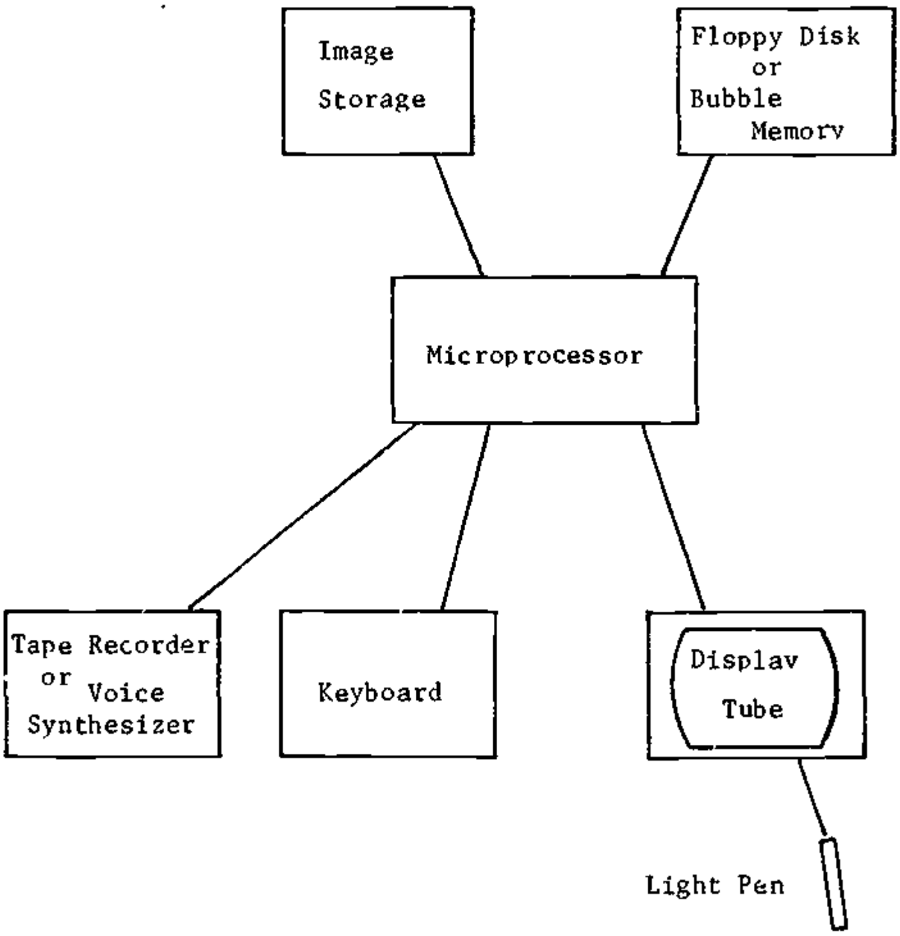


Image Storage
Floppy Disk or Bubble Memorv
Microprocessor
Tape Recorder or Voice Synthesizer
Keyboard
Displav Tube
Light Pen

F.gure 5.3d A Microprocessor-Based Testing System

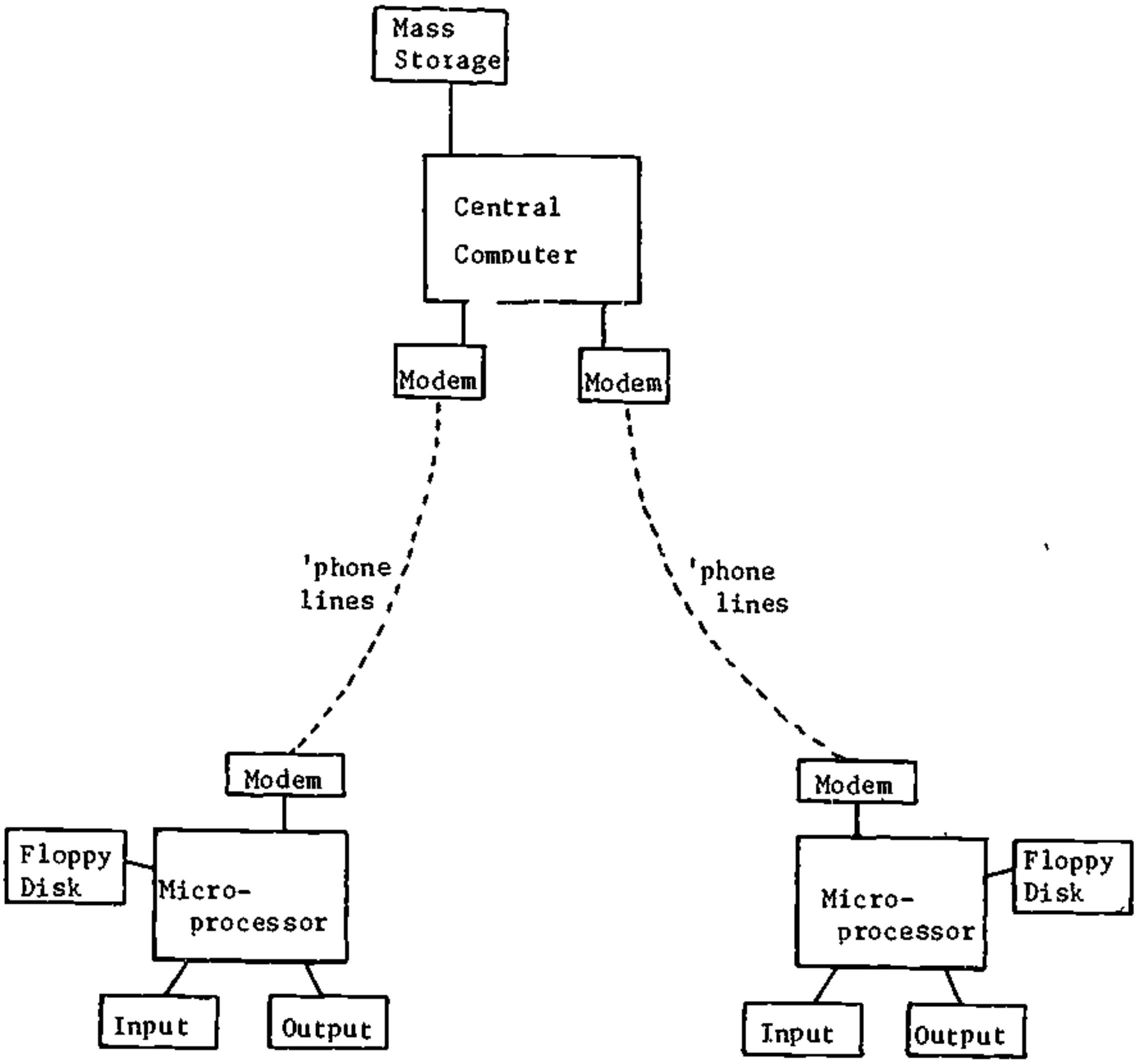


Figure 5.3e A Hybrid Network

retrieving graphic images at reasonable cost. Recently, however, image processing devices have been designed that permit the storage of graphic material including (in some cases) color and motion. Various image processing technologies are available. The plasma display technology employed in the PLATO system is one example. Video discs, because of their large storage capacity, random access, and ability to produce motion, appear to have significant potential in testing environments. Recent advances in microform technology may also provide inexpensive random access to graphic data.

5.4 Recommendation 3: That extensions to item response theory and the development of alternative models for the provision of adaptive testing be explored. Such models might include item selection strategies especially suited to microprocessors; multidimensional trait models; models for achievement testing; and models for use in diagnostic and mastery learning environments in which items are linked to learning objectives. A related area of importance is the construction of adaptive test batteries in which branching occurs at the test level as well as the item level.

As was pointed out in Chapter 1, the Broad-Range Tailored Test implements one of a number of designs which might have been chosen for adaptive testing. Among alternatives to the maximum-likelihood estimation procedure are Bayesian estimation procedures and the Weiss (1973) stradaptive procedure. Even within the context of the maximum likelihood estimation alternative stopping rules, different choices for the initial item, variations in step size, and alternative structures for the item pool could have been employed. There are many research issues which deserve exploration. Although a few will be mentioned in this document, it is recommended that the input of interested staff be solicited regarding future developments in this area.

One area of interest is the development of multidimensional models. The unidimensionality assumption of latent trait theory has been taken to be a serious constraint by some researchers; this is particularly true in the domain of achievement testing in which the test typically involves a multidimensional space. Sympson (1977) has developed a multidimensional latent trait model for dichotomously scored multiple-choice items. Urry (1977) has developed a multidimensional Bayesian approach to tailored testing. Many questions remain to be answered about the appropriateness of multidimensional models and of the most effective computational techniques for their use.

Another area of potential interest is the development of item selection strategies especially suited to microprocessors. Jones (1979, unpublished manuscript) has argued that strategies based on sequential analysis may be pressed into service for use in adaptive testing with considerable savings in computational time.

Most of the work in adaptive testing has involved the estimation of ability. In a mastery learning environment the estimation of achievement is generally far more important. Research needs to be conducted into the development of adaptive testing models which can be used for measuring achievement, especially models which link items to learning objectives. It is possible that such models could employ item response theory in unidimensional or multidimensional models. However, it would be useful to explore non-trait models. The concept that individuals possess fixed traits has been challenged by such social learning theorists as Bandura and Mischel, who argue for alternative constructs such as response tendency and state. True score models may not be consistent with a social learning perspective which stresses situational variables. Social learning models place an increased emphasis on individual differences. All too often individual differences have been considered "error," even though individual variation may be a valuable important source of information about a person. Since (unlike a paper-and-pencil instrument), the computer has the capability of adapting to individual differences, individual difference models may play a key role in computerized testing.

Tests which yield detailed individualized profiles for diagnosis and educational prescription are desirable. Unfortunately, comprehensive profiles are difficult to construct through conventional tests because of the large number of items needed to obtain reliable scores. Adaptive testing will be useful in areas where comprehensive individual profiles are desired. Thus, if a dichotomous classification (select vs. reject, remedial vs. standard) were to be made on the basis of a preestablished "cut score," a conventional test might be most appropriate. However, if the purpose of the test were to design an individualized educational program or to place an individual in an appropriate vocational training program, an adaptive test would tend to be more effective. In developing profiles, the computer can present batteries of adaptive tests in which branching would take place among subtests as well as among items. As with single tests, adaptive batteries can be conceptualized in both unidimensional and multidimentional form.

Because of the multiplicity of the research issues and the limitations of the resources available, Recommendation 1 suggests that research priorities be guided by development priorities. Much of the research proposed can take place in the context of developing adaptive tests for use outside the laboratory. Some laboratory research is desirable because of the greater possibilities for experimental control; thus, analytic and simulation models may be cost-effective techniques for model development in the early stages. It should also be noted that adaptive testing research can provide useful insights into the construction of paper-and-pencil tests. There is little reason for isolating adaptive testing research from the mainstream of psychometric research since there is much to be gained from its integration.

5.5 Recommendation 4: That high priority be accorded to the development of innovative item types for computer presentation. Such items might involve simulation and gaming, constructed responses, graphics, motion, sound, and time-dependent responses

The objective multiple-choice item has been the mainstay of testing for many years. So entrenched is this format that some people might be tempted to conclude it is the most desirable. In fact, the multiple-choice objective item has many advantages, including its standardization, the limited range of response possibilities it allows, its fit with psychometric models, and the ease with which it may be scored. However, in many respects the objective multiple-choice item is an artifact of the economics involved in mass testing of large numbers of persons. As a measurement option, the multiple-choice item has a number of limitations. One limitation is that the major cognitive process measured by the multiple choice item is the individual's ability to discriminate a correct response from among a series of alternatives. It is well established that the psychological processes involved in recognition are different from other important processes such as recall, synthesis, and evaluation which the objective multiple-choice item can only measure indirectly. Because the multiple-choice item cannot readily measure divergent responses, it is limited in its ability to assess problem solving.

The computer has the potential to free the test developer from the constraints imposed by a multiple-choice format. Many novel item formats are possible. Items presented by computer may employ constructed responses. Probabilistic response models in which an individual weights different alternatives may be employed. Items may be constructed that require time-dependent responses; for example, an individual may be asked to listen to a conversation and press a button when a grammatical error occurs. Items may employ graphics; for example, mathematical concepts may be tested by asking the individusl's being asked to group objects, draw lines, or construct angles. Process conceptions may be tested by our asking the test taker to follow the flow of a process diagram with a light pen. The possibilities of computer-administered items that have the potential to support novel forms of testing go far beyond the capabilities of the objective multiple-choice item.

It is recommended that the organization accord high priority to the exploration of computer-presented items because new techniques may form the basis for novel assessment procedures. Weiss (1977) has predicted that "the multiple-choice item will disappear and exist only in museums. We will learn how to use graded responses, continuous responses, and free responses; and in the process we will humanize testing even a little bit more, not only by adapting the test to individual differences and abilities and other variables, but also by allowing people to respond in a more natural way than is allowed by multiple choice tests."

As discussed in Recommendation 1, the development of creative solutions requires synergy among experts in test development, computer technology, psychometrics, and education. Few test development professionals have sufficient knowledge of capabilities of computers to evaluate the potential of this technology. Technically oriented persons rarely understand the subtleties of test development. Psychometric support is needed to develop models for scoring and interpretation of novel assessment procedures and to insure that new techniques developed rest on a firm theoretical foundation. Finally, it is important that psychologists and educational specialists be involved so that the assessment techniques developed bear a relationship to real problems and real people. The importance of a multidisciplinary approach cannot be overemphasized.

Chapter VI

Preliminary Study

Before conducting the main study, it was necessary to investigate the effectiveness of the human/computer interface. Accordingly, a pilot study was conducted:

1. To determine if students of various levels and ages could interact with the computer system easily and without confusion.

2. To determine if modifications to the system protocols would yield a more effective human/computer interface.

3. To determine if the length of time required to administer the BRTT and the extent to which fatigue factors would affect performance when both forms were administered.

4. To determine if practice effects would systematically bias the relationship between the first and second form administered.

5. To refine the posttest questionnaire on the basis of student feedback.

6. To refine the instruction given to the students.

Students participating in the preliminary study were 5th grade students (N=3), 7th grade students (N=3), high school sophomores (N=11), and adults (N=6).

The high school students, the 7th grade students, and three of the adults were given both forms of the BRTT. The remainder of the subjects were given a single form of the BRTT. The time required to take each form of the BRTT was noted as were technical difficulties, requests for help, and apparent ease of the human/computer protocol. Following the tests, the subjects were asked their opinions of the experience. The high school students were asked to respond to a formal set of questions and to participate in a unstructured discussion group in which reactions to the testing experience were discussed.

The subjects were able to complete the tests in a relatively short time. The mean time for administration of Form A of the BRTT was 18 minutes; for Form B it was 18.2 minutes. Only one subject (a fifth grader) required more than 25 minutes to complete the test. For the 11 high school students, ability estimates ranged from .10 to 1.83; difference scores ranged from a low of .08 to a high of .68. The test-retest correlation for the high school students was .53.

The students generally felt that using the terminal rather than a paper-and-pencil test was more enjoyable. They felt the terminal to be less fatiguing and generally a beneficial experience. More than half of the students said that they were relaxed during the testing situation. It was observed that most students preferred the idea that answer sheets, test booklets, etc, were not necessary.

Some students noted that they felt pressured whenever someone completed the test ahead of them. (Students were assured, however, that this was not a timed test). Others mentioned that the "buzzer" that sounded when an error was made was startling.

A common complaint was that the letter "g" on the DEC VT52 terminal was hard to read. Unfortunately, this was a hardware function and could not be changed.

One case of hardware failure occurred. This was determined to be due to static electricity resulting from a chair's friction with the carpet.

Following are the high school students' responses to the posttest questionnaire.

Table 6.1a

Responses of High School Students to Posttest Questions

Item	Response	#	%
A. Previous Experience with Computers			
1. Are you at all familiar with computers?	Yes	2	1B
	No	9	B2
2. Have you ever punched computer cards at a keypunch machine before?	Yes	4	36
	No	7	64
3. Have you ever interacted with a computer by means of a terminal before?	Yes	3	27
	No	8	73
B. Perception of Test Difficulty			
4. How often did you feel that the questions in the test were too easy for you?	Always	0	0
	Frequently	0	0
	Sometimes	6	54
	Seldom	3	27
	Never	2	18
5. How often did you feel that the questions in the test were too hard for you?	Always	0	0
	Frequently	4	36
	Sometimes	7	64
	Seldom	0	0
	Never	0	0
6. On how many of the questions did you guess?	Almost all	0	0
	More than half	0	0
	About half	2	18
	Less than half	6	54
	Almost none	3	27
7. How often were you sure that your answer to the questions were correct?	Almost always	0	0
	More than half	3	27
	About half	6	54
	Less than half	2	18
	Almost never	0	0

Table 6.1a (continued)

Item	Response	#	%
8. In relation to your vocabulary ability, how difficult was the test for you?	Too difficult	0	0
	Somewhat difficult	9	82
	About right	2	18
	Somewhat east	0	0
	Too easy	0	0
9. Did you feel frustrated by the difficulty of the test questions?	Not at all	3	27
	Somewhat	7	64
	Fairly much so	1	9
	Very much so	0	0
C. Anxiety and Motivation	Response	#	%
10. During testing, did you worry about how well you would do?	Not at all	1	9
	Somewhat	8	73
	Fairly much so	2	18
	Very much	0	0
11. Were you nervous while taking the test?	Not at all	9	82
	Somewhat	1	9
	Moderately so	1	9
	Very much so	0	0
12. How did you feel while taking the test?	Very tense	0	0
	Somewhat tense	2	18
	Neutral	1	9
	Somewhat relaxed	5	45
	Very relaxed	3	27
13. Did nervousness while taking the test prevent you from doing your best?	Difinitely	0	0
	Somewhat	1	9
	Probably not	6	54
	Definitely not	4	36
14. How frequently were you careful to select what you thought was the best answer to each question?	Almost always	6	54
	Frequently	5	45
	Sometimes	0	0
	Rarely	0	0
	Never	0	0

Table 6.1a (continued)

Item	Response	#	%
15. Did you feel challenged to do as well as you could on the test?	Not at all	0	0
	Somewhat	3	27
	Fairly much	5	45
	Very much	3	27
16. Did you care how well you did on the test?	Yes, a lot	3	27
	Yes, a little	8	73
	A little	0	0
	Very little	0	0
	Not at all	0	0
D. Factors Related to Computer Administration			
17. Did the mechanics of using the computer terminal interfere with your taking the test?	Not at all	8	73
	Slightly	3	27
	Somewhat	0	0
	Very much	0	0
18. How tiring did you find the computer-administered test?	Very	0	0
	Somewhat	1	9
	Slightly	4	36
	Not at all	6	54
19. Compared to a "paper-and-pencil" multiple-choice test of the same length would you:			
Find the computer test more tiring?		0	0
Find both tests about the same?		1	9
Find the paper-and-pencil test more tiring?		10	91
20. Which would you prefer to take?			
A computer-administered test		9	82
A pencil-and-paper test		0	0
No preference		2	18

Table 6.1 a (continued)

Item	#	%
E. Desirability of Item Feedback		
The computer could score each item as you answer it and tell you if your choice was right or wrong. This is called <u>feedback</u>.		
21. Would getting feedback on the test make it:		
More interecting	8	73
Less interesting	1	9
Cannot say	2	18
22. Would getting feedback after each question make you nervous?		
Very nervous	1	9
Somewhat nervous	2	18
Slightly nervous	5	45
Not nervous at all	3	27
23. How would you feel about getting feedback?		
I would rather not know whether my answers were right or wrong.	3	27
I really don't care if I got feedback or not.	1	9
I would like getting feedback.	7	64

Appendix A

Posttest Questionnaire

The posttest questionnaire was designed to obtain demographic data regarding the student population and attitudinal data about student reaction to the adaptive test. The results of the questionnaire are reported in Chapter IV. Responses to the pilot version of the questionnaire will be found in Chapter VI, which describes the preliminary study. This appendix contains the questionnaire in its original form. The attitude variables assessed by the questionnaire are: prior familiarity with computers, subjective perception of difficulty, anxiety, motivation, human factors in the computer/human interface, preference for adaptive vs. conventional testing, and feedback. To facilitate comparison with previous research, the items on the questionnaire were adapted from items used by previous investigators.

Most of the questions were of the objective multiple-choice type. Three of the items (37-40) employed an incomplete sentence blank format that permitted free response.

Please provide for us the following background information. Your responses will be kept strictly confidential. If you strongly object to answering any question please feel free to omit it.

Name ____

Date of Birth ______________________________________

Your Sex ________ Male ________ Female

High School Attended _______________________________

Years in School ____________________________________

Have you taken the PSAT? _________ Yes _________ No

If yes, when did you take it? ______________________

Have you taken the SAT? _________ Yes _________ No

If yes, when did you take it? ______________________

Are you planning to take the SAT at a future time? _________ Yes _________ No

If yes, when _______________________________________

What is your High School.average? __________________

What is the highest possible average at your school?_______________

What are your plans when you graduate from High School?

__________ attend 4 year college

__________ attend 2 year college

__________ work

__________ other ___________________________________

Are you at all familiar with computers? ________ Yes ________ No

Have you ever punched computer cards at a keypunch machine before?

________ Yes ________ No

Have you ever interacted with a computer by means of a terminal before?

__________ Yes ________ No

How often did you feel that the questions in the test were too easy for you?

______ a. Always

______ b. Frequently

______ c. Sometimes

______ d. Seldom

______ e. Never

How often did you feel that the questions in the test were too hard for you?

______ a. Always

______ b. Frequently

______ c. Sometimes

______ d. Seldom

______ e. Never

On how many of the questions did you guess?

______ a. Almost all of the questions

______ b. More than half of the questions

______ c. About half the questions

______ d. Less than half of the questions

______ e. Almost none of the questions or never

How often were you sure that your answers to the questions were correct?

______ a. Almost always

______ b. More than half of the time

______ c. About half of the time

______ d. Less than half of the time

______ e. Almost never

In relation to your vocabulary ability, how difficult was the test for you?

________ a. Much too difficult

________ b. Somewhat too difficult

________ c. Just about right

________ d. Somewhat too easy

________ e. Much too easy

Did you feel frustrated by the difficulty of the test questions?

________ a. Not at all

________ b. Somewhat

________ c. Fairly much so

________ d. Very much so

During testing, did you worry about how well you would do?

________ a. Not at all

________ b. Somewhat

________ c. Fairly much so

________ d. Very much

Were you nervous while taking the test?

________ a. Not at all

________ b. Somewhat

________ c. Moderately so

________ d. Very much so

How did you feel while taking the test?

________ a. Very tense

________ b. Somewhat tense

________ c. Neither tense nor relaxed

________ d. Somewhat relaxed

________ e. Very relaxed

Did nervousness while taking the test prevent you from doing your best?

________ a. Yes, definitely

________ b. Yes, somewhat

________ c. Probably not

________ d. Definitely not

How frequently were you careful to select what you thought was the best answer to each question?

________ a. Almost always

________ b. Frequently

________ c. Sometimes

________ d. Rarely

________ e. Never

Did you feel challenged to do as well as you could on the test?

________ a. Not at all

________ b. Somewhat

________ c. Fairly much so

________ d. Very much so

Did you care how well you did on the test?

________ a. I cared a lot

________ b. I cared some

________ c. I cared a little

________ d. I cared very little

________ e. I didn't care at all

Did the mechanics of using the computer terminal interfere with your taking the test:

________ a. Not at all

________ b. Slightly

________ c. Somewhat

________ d. Very much so

How tiring did you find the computer-administered test?

_______ a. Very tiring

_______ b. Somewhat tiring

_______ c. Slightly tiring

_______ d. Not tiring at all

Compared to a "paper-and-pencil" multiple-choice test of the same length would you

_______ a. Find the computer test more tiring?

_______ b. Find both tests about the same?

_______ c. Find the paper-and-pencil test more tiring?

Which would you prefer to take?

_______ a. A computer-administered test

_______ b. A pencil-and-paper test

_______ c. No preference

The computer could score each item as you answer it and tell you if your choice was right or wrong. This is called feedback.

Would getting feedback on the test make it:

_______ a. More interesting

_______ b. Less interesting

_______ c. Cannot say

Would getting feedback after each question make you nervous?

_______ a. Very nervous

_______ b. Somewhat nervous

_______ c. Slightly nervous

_______ d. Not nervous at all

How would you feel about getting feedback?

______ a. I would rather not know whether my answers were right or wrong.

______ b. I really don't care if I got feedback or not.

______ c. I would like getting feedback.

What did you like best about the computer-administered test? __________

____

____

____

What did you like least? ______________________________________

____

____

____

How could we change the test to improve it? ____________________

____

____

____

____

Appendix B

Description of the Item Pool

SUMMARY SHEET

ITEM TYPE

# 1 - Synonyms

2 - Opposites

3 - Incomplete Sentences

4 - Word Relations

5 - Sentence Comprehension

SCAT I (form 2A, 2B, 3A, 3B, 4A) contributed 65 items. The items consisted of synonyms that used instruction set (1), and incomplete sentences that used instruction set (5).

SCAT II (form 1A, 2A, 2B, 3A, 3B, 4A) contributed 107 items. All items were word relations which utilized instruction set (9).

STEP II contributed 39 items. All items were sentence comprehension. Instruction set (13) was used.

PSAT contributed 56 items. Opposites, incomplete sentences, and word relations were the various types that were employed. Instruction set (3) was used with items that were opposites, (7) with incomplete sentences, and (11) with word relations.

SAT contributed 27 items. Type 2, 3, and 4 were used. Instruction set (3), (7), and (11) were used respectively.

GRE contributed 69 items. Type 2, 3, and 4. Instruction set (3), (7), and (11) were used respectively.

Item Source Test & Form Name	Item#	Record Number	Instruction Set	Answer Key	IPA	IPB	IPC	Item Type	Test Form Code	# of Items Contributec	Line#	INDEX
SCAT I 3A	37	351	1	A	1.51	-.42	.16	1	53	22	2-6	13A
	42	352	1	B	1.91	-.35	.25	1	53	22	2-6	
	22	353	1	A	1.60	-.98	.14	1	53	22	2-6	
	13	354	1	E	.71	-1.28	.14	1	53	22	2-6	
	1	355	1	E	.51	-3.81	.14	1	53	22	2-6	
	4	356	1	D	.73	-2.46	.14	1	53	22	2-6	
	3	357	1	C	.90	-2.74	.14	1	53	22	2-6	
	279	358	5	D	1.56	.18	.08	3	53	22	5-9	
	304	359	5	B	1.56	.63	.15	3	53	22	3-7	
	308	360	5	E	.92	.69	.14	3	53	22	4-8	
	234	361	5	E	.72	-.68	.14	3	53	22	3-7	
	254	362	5	A	1.09	-.26	.14	3	53	22	5-9	
	238	363	5	D	1.0	-.54	.14	3	53	22	4-8	
	226	364	5	B	1.58	-.92	.10	3	53	22	4-8	
	205	365	5	E	.77	-1.99	.14	3	53	22	4-8	
SCAT I 3B	59	301	1	D	1.08	-.02	.15	1	54	17	2-6	13B
	47	302	1	E	1.67	-.27	.19	1	54	17	2-6	
	14	303	1	D	1.28	-1.16	.15	1	54	17	2-6	
	6	304	1	A	.60	-2.07	.15	1	54	17	2-6	
	11	305	1	E	1.34	-1.56	.15	1	54	17	2-6	
	26	306	1	B	1.30	-.91	.15	1	54	17	2-6	
	5	307	1	B	.76	-2.26	.15	1	54	17	2-6	
	12	308	1	E	1.28	-1.39	.15	1	54	17	2-6	
	2	309	1	A	.59	-2.83	.15	1	54	17	2-6	
	358	310	5	D	1.27	1.42	.10	3	54	17	4-8	
	246	311	5	B	.86	-.38	.15	3	54	17	4-8	
	269	312	5	E	.45	.07	.15	3	54	17	3-7	
	258	313	5	E	1.50	-.22	.10	3	54	17	3-7	
	250	314	5	B	1.63	-.32	.15	3	54	17	3-7	
	213	315	5	B	1.36	-1.49	.15	3	54	17	4-8	
	209	316	5	B	.66	-1.88	.15	3	54	17	4-8	
	207	317	5	C	1.33	-1.93	.15	3	54	17	3-7	

Item Source Test & Form Name	Item#	Record Number	Instruction Set	Answer Key	IPA	IPB	IPC	Item Type	Test Form Code	# of Items Contributed	Line#	INDEX
SCAT II IA	748	135	9	C	1.48	2.97	.11	4	55	19	2-5	II IA
	757	136	9	B	.84	3.85	.31	4	55	19	2-5	
	741	137	9	C	1.48	2.51	.17	4	55	19	2-5	
	737	138	9	A	.67	2.40	.09	4	55	19	2-5	
	714	139	9	C	.66	1.91	.20	4	55	19	2-5	
	729	140	9	C	1.35	2.23	.24	4	55	19	2-5	
	747	141	9	B	1.48	2.85	.28	4	55	19	2-5	
	743	142	9	A	.77	2.58	.23	4	55	19	2-5	
	716	143	9	A	1.04	1.93	.17	4	55	19	2-5	
	723	144	9	D	.62	2.06	.20	4	55	19	2-5	
	684	145	9	D	.82	1.26	.09	4	55	19	2-5	
	682	146	9	B	.89	1.23	.09	4	55	19	2-5	
	667	147	9	A	.84	1.03	.17	4	55	19	2-5	
	655	148	9	B	.94	.82	.09	4	55	19	2-5	
	606	149	9	A	.57	.28	.17	4	55	19	2-5	
	679	150	9	B	.81	1.20	.17	4	55	19	2-5	
	472	151	9	D	.72	-1.47	.17	4	55	19	2-5	
	570	152	9	A	1.04	-.05	.17	4	55	19	2-5	
	484	153	9	A	.72	-1.29	.17	4	55	19	2-5	
SCAT II 2A	692	154	9	B	1.74	1.43	.06	4	56	9	2-5	II 2A
	711	155	9	C	1.60	1.81	.17	4	56	9	2-5	
	696	156	9	D	1.07	1.52	.20	4	56	9	2-5	
	688	157	9	B	.92	1.34	.20	4	56	9	2-5	
	672	158	9	A	.69	1.12	.22	4	56	9	2-5	
	650	159	9	C	.84	.74	.24	4	56	9	2-5	
	615	160	9	A	.52	.41	.20	4	56	9	2-5	
	560	161	9	B	.70	-.15	.24	4	56	9	2-5	
	459	162	9	A	.75	-1.68	.20	4	56	9	2-5	
SCAT II 2B	740	163	9	D	1.36	2.45	.12	4	57	15	2-5	
	738	164	9	A	.75	2.41	.15	4	57	15	2-5	
	706	165	9	D	1.26	1.65	.27	4	57	15	2-5	
	700	166	9	D	1.05	1.60	.11	4	57	15	2-5	
	674	167	9	B	1.20	1.17	.20	4	57	15	2-5	

STEP II 3B	865	262	13	B	1.08	- .05	.16	5	64	14	5-8	II 3B
	870	263		D	1.45	.12	.20				7-10	
	856	264		B	1.62	- .27	.31				6-9	
	806	265		C	1.26	-1.65	.21				5-8	
	824	266		C	1.15	-1.07	.20					
	859	267		B	.92	- .20	.23				4-7	
	815	268		B	1.36	-1.34	.20				5-8	
	780	269		B	.63	-2.66	.20				4-7	
	809	270		A	.97	-1.50	.20				5-8	
	772	271		C	.75	-3.19	.20					
	791	272		B	.62	-2.06	.20					
	794	273		A	.69	-1.91	.20				4-7	
	779	274		B	.80	-2.72	.20				5-8	
	804	275		A	1.00	-1.68	.20				6-9	
STEP II 4A												
	868	242		C	1.39	.06	.22		65	20	5-8	II 4A
	844	243		D	1.05	- .68	.22					
	785	244		C	.97	-2.44	.27					
	773	245		C	.94	-3.19	.15				6-9	
	781	246		D	.81	-2.54	.2					
	784	247		A	.82	-2.46	.24				5-8	
	774	248		A	.98	-3.13	.2				4-7	
	771	249		D	.79	-3.42	.2					
	776	250		C	.90	-3.01	.23					
	778	251		A	.66	-2.78	.2				5-8	
	782	252		B	.64	-2.52	.24				4-7	
	770	253		C	1.16	-3.57	.2				5-8	
	765	254		D	.66	-4.13	.2					
	767	255		C	1.29	-3.87	.2					
	769	256		B	.92	-3.60	.2					
	768	257		D	.78	-3.83	.15					
	766	258		D	.78	-3.87	.2				4-7	
	764	259		C	.70	-4.40	.2				6-9	
	762	260		C	.78	-4.79	.2					
	763	261		A	.58	-4.73	.2				5-8	

Item Source Test and Form Name	Item #	Record #	Instr. Set #	Answer Key	IPA	IPB	IPC	Item Type	Test Form Code	# Items Cont	Line #	INDEX
SCAT II 4A	413	236	9	C	.32	-5.26	.20	4	60	35	2-5	II 4A
	415	237	9	D	.76	-4.83	.20	4	60		2-5	
	426	238	9	B	.45	-3.00	.20	4	60		2-5	
	412	239	9	C	.60	-5.45	.20	4	60		2-5	
	410	240	9	B	.54	-5.53	.20	4	60		2-5	
	411	241	9	C	.51	-5.48	.20	4	60		2-5	
STEP II 2A												
	893	293	13	D	.57	1.11	.15	5	61	8	5-8	II 2 A
	897	294		B	1.12	1.53	.32				6-9	
	874	295		A	1.74	.28	.24					
	879	296		B	.89	.60	.24				5-8	
	848	297		D	1.14	- .60	.15					
	840	298		A	1.12	- .74	.15					
	799	299		C	.61	-1.81	. 2				4-7	
	828	300		A	1.11	- .98	.24				5-8	
STEP II 2B												
	891	287	13	A	.99	.91	. 2	5	62	6	5-8	II 2B
	887	288		A	1.75	.85	. 2					
	872	289		D	.83	.15	. 2					
	853	290		B	.76	- .40	. 2				6-9	
	832	291		B	.79	- .94	. 2				5-8	
	821	292		A	1.48	-1.10	.15				6-9	
STEP III 3A												
	883	276	13	D	1.50	.68	.16	5	63	11	4-7	II 3A
	862	277		D	.83	- .14	.2				7-10	
	836	278		C	1.50	- .77	.2				6-9	
	812	279		A	.59	-1.38	.2				5-8	
	796	280		B	.56	-1.85	.2				6-9	
	818	281		D	1.23	-1.29	.24				5-8	
	801	282		C	.83	-1.75	.2					
	786	283		C	.87	-2.41	.2					
	783	284		D	.87	-2.48	.2					
	777	285		C	.99	-2.80	.2				4-7	
	775	286		D	.73	-3.10					5-8	

PSAT UPT 1	154	111	3	B	2.21	1.72	.20	2	66	16	2-6	UI
	173	112		D	2.47	2.29	.25					
	139	113		D	.86	.98	.15					
	327	114	7	E	.94	.99	.15	3			3-7	
	300	115	7	D	.85	.56	.15	3			4-8	
	331	116		A	1.44	1.09	.15				56-10	
	366	117		C	1.30	1.57	.12				4-8	
	356	118		A	.80	1.37	.15					
	362	119		A	.65	1.53	.1				6-10	
	347	120		D	1.99	1.26	.1					
	379	121		D	1.89	1.95	.19				4-8	
	385	122		C	1.05	2.30	.15				5-9	
	386	123		C	3.03	2.32	.16				4-8	
	611	124		A	1.38	.34	.15	4			2-6	
	690	125		D	1.68	1.40	.15	4				
	731	126		C	2.96	2.28	.16	4				
PSAT UPT Z												
	150	127		C	1.19	1.45	.15	2	67	8	2-6	UZ
	145	128		B	1.64	1.22	.10					
	316	129		E	.88	.83	.15	3			6-10	
	372	130		B	1.56	1.76	.06				4-8	
	377	131		B	2.96	1.89	.07					
	397	132		E	1.70	3.07	.10				5-9	
	694	133		E	.98	1.48	.10	4			2-6	
	733	134		E	.76	2.32	.10	4				
PSAT QPT 1												
	312	99	7	B	.87	.75	.15	3	68	12	3-7	Q
	364	100		D	1.34	1.54	.19				5-9	
	374	101		E	1.89	1.84	.1					
	380	102		C	.99	2.08	.1					
	383	103		C	1.45	2.22	.13				3-7	
	395	104		D	1.01	2.90	.12				4-8	
	628	105		C	1.22	.54	.1	4			2-6	
	588	106		B	1.43	.12	.15					
	659	107		D	1.21	.86	.04					
	669	108		D	1.16	1.07	.15					

Item Source Test & Form Name	Item#				IPA	IPB	IPC	Item Type	Test form Code		Line#	INDEX
SCAT II 2B	619	168	9	D	1.00	.44	.15	4	57	15	2-5	II 2B
	623	169	9	C	.55	.48	.20	4	57	15	2-5	
	646	170	9	C	.65	.71	.27	4	57	15	2-5	
	539	171	9	A	1.25	-.50	.20	4	57	15	2-5	
	546	172	9	D	.86	-.36	.20	4	57	15	2-5	
	468	173	9	C	.71	-1.51	.20	4	57	15	2-5	
	503	174	9	C	.51	-1.03	.20	4	57	15	2-5	
	553	175	9	C	.80	-.28	.20	4	57	15	2-5	
	480	176	9	A	.46	-1.38	.20	4	57	15	2-5	
	496	177	9	D	.87	-1.14	.20	4	57	15	2-5	
SCAT II 3A	584	178	9	D	.96	.07	.17	4	58	14	2-5	II 3A
	567	179	9	D	.95	-.08	.15	4	58	14	2-5	
	556	180	9	C	.99	-.19	.24	4	58	14	2-5	
	445	181	9	A	.41	-1.78	.20	4	58	14	2-5	
	524	182	9	B	.48	-.69	.20	4	58	14	2-5	
	457	183	9	C	1.10	-1.73	.15	4	58	14	2-5	
	445	184	9	C	.58	-1.97	.20	4	58	14	2-5	
	432	185	9	D	.36	-2.55	.20	4	58	14	2-5	
	436	186	9	A	.81	-2.44	.20	4	58	14	2-5	
	534	187	9	A	.56	-.52	.15	4	58	14	2-5	
	527	188	9	B	.49	-.64	.15	4	58	14	2-5	
	447	189	9	A	.85	-1.94	.20	4	58	14	2-5	
	434	190	9	D	1.03	-2.51	.20	4	58	14	2-5	
	549	191	9	A	1.15	-.34	.29	4	58	14	2-5	
SCAT II 3B	664	192	9	A	1.43	.91	.15	4	59	15	2-5	II 3B
	641	193	9	D	1.23	.65	.20	4	59	15	2-5	
	602	194	9	C	.94	.27	.20	4	59	15	2-5	
	637	195	9	B	1.21	.63	.20	4	59	15	2-5	
	593	196	9	D	1.09	.19	.24	4	59	15	2-5	
	510	197	9	B	.63	-.83	.20	4	59	15	2-5	
	575	198	9	C	1.13	-.00	.27	4	59	15	2-5	
	632	199	9	C	1.27	.57	.22	4	59	15	2-5	
	517	200	9	C	1.01	-.75	.21	4	59	15	2-5	
	453	201	9	A	.46	-1.82	.20	4	59	15	2-5	

Test & Form Name	Item#	Record Number	Instruction Key	Answer Key	IPA	IPC	IPC	Item Type	Test Form Code	# of Item Contribution	Line#	INDEX
SCAT II 3B	441	202	9	C	.57	-2.19	.20	4	59	15	2-5	II 3B
	448	203	9	A	.80	-1.22	.15	4	59	15	2-5	
	430	204	9	C	.86	-2.79	.20	4	59	15	2-5	
	433	206	9	B	.77	-2.55	.20	4	59	15	2-5	
	419	206	9	B	.45	-3.96	.20	4	59	15	2-5	
SCAT II 4A	542	207	9	B	.76	-.45	.18	4	60	35	2-5	II 4A
	520	208	9	D	1.11	-.73	.15	4	60		2-5	
	579	209	9	A	1.39	.01	.22	4	60		2-5	
	464	210	9	D	.76	-1.5	.20	4	60		2-5	
	492	211	9	B	1.39	-1.18	.20	4	60		2-5	
	506	212	9	A	.80	-.98	.20	4	60		2-5	
	500	213	9	A	.65	-1.07	.15	4	60		2-5	
	449	214	9	B	1.27	-1.88	.20	4	60		2-5	
	476	215	9	D	.80	-1.45	.24	4	60		2-5	
	531	216	9	C	1.39	-.55	.30	4	60		2-5	
	513	217	9	D	.87	-.83	.33	4	60		2-5	
	451	218	9	C	.75	-1.88	.20	4	60		2-5	
	431	219	9	D	.63	-2.79	.20	4	60		2-5	
	435	220	9	C	.59	-2.49	.20	4	60		2-5	
	427	221	9	C	.89	-2.99	.20	4	60		2-5	
	443	222	9	A	.74	-2.07	.15	4	60		2-5	
	421	223	9	A	.60	-3.49	.20	4	60		2-5	
	428	224	9	B	.98	-2.87	.24	4	60		2-5	
	423	225	9	A	.73	-3.39	.15	4	60		2-5	
	425	226	9	D	.70	-3.07	.20	4	60		2-5	
	437	227	9	B	.47	-2.43	.20	4	60		2-5	
	416	228	9	A	.40	-4.22	.20	4	60		2-5	
	420	229	9	C	.72	-3.53	.20	4	60		2-5	
	424	230	9	A	.54	-3.37	.15	4	60		2-5	
	429	231	9	B	1.09	-2.83	.24	4	60		2-5	
	422	232	9	A	.62	-3.43	.20	4	60		2-5	
	417	233	9	A	.47	-4.17	.20	4	60		2-5	
	414	234	9	B	.54	-4.91	.20	4	60		2-5	
	418	235	9	D	.53	-4.14	.20	4	60		2-5	

Item Source	Item #	Rec. #	Instr. Set	Ans.	IPA	IPB	IPC	Type	Form	Items Cont.	Line #	Index
GRE KQGR 2	323	44	7	C	.64	.93	.19	3	72		4-8	G2
	343	45		A	.97	1.16	.19					
	264	46		A	.66	.00	.19				3-7	
	352	47		B	.72	1.31	19					
	339	48		D	.63	1.13	.19				4-8	
	402	49		E	1.00	3.40	.19					
	391	50		C	.72	2.70	.18				5-9	
	720	51	11	C	.77	1.97	.09	4			2-6	
	698	52	11	E	.69	1.56	.19					
	745	53		D	1.12	2.66	.19					
	751	54		E	.65	3.15	.16					
	759	55		B	1.11	4.25	.29					
	755	56		B	11.04	3.70	.19					
	760	57		D	.87	4.50	.16					
	761	58		E	.57	4.71	.19					
	167	59	3	B	.60	2.20	.19	2				
	176	60		C	.61	2.34	.19					
	193	61		D	1.40	3.71	.34					
	201	62		E	.52	4.69	.25					
	390	63	7	D	.70	2.67	.18	3			3-7	
	399	64		E	.81	3.13	.19				4-8	
	400	65		D	.93	3.22	.19				5-9	
	388	66		C	.56	2.53	.07				4-8	
	403	67		D	1.33	3.45	.22					
	406	68		E	1.40	3.74	.16				5-9	
	404	69		B	.98	3.56	.27				3-7	
	408	70		C	.91	4.38	.16				5-9	
	409	71		C	1.24	4.60	.08					

Item Source Test & Form Name	Item#	Record#	Instr. Set No.	Answer Key	Discrimination IPA	Difficulty IPB	Guessing IPC	Item Type	Test Form Code	# of Items Contributed	Line#	INI
SCAT I 2A	108	318	1	C	1.78	1.37	.18	1	51	15	2-6	12
	101	319	1	D	1.86	.98	.13	1	51	15	2-6	
	113	320	1	B	1.86	1.90	.14	1	51	15	2-6	
	77	321	1	A	1.86	.38	.19	1	51	15	2-6	
	10	322	1	A	1.86	.38	.19	1	51	15	2-6	
	18	323	1	B	1.17	1.01	.11	1	51	15	2-6	
	8	324	1	A	.90	1.80	.13	1	51	15	2-6	
	30	325	1	C	.42	-.82	.13	1	51	15	2-6	
	375	326	5	D	1.03	1.88	.13	3	51	15	6-10	
	360	327	5	D	.81	1.45	.11	3	51	15	4-8	
	274	328	5	D	.87	.15	.06	3	51	15	4-8	
	284	329	5	D	1.17	.35	.13	3	51	15	3-7	
	242	330	5	A	1.09	-.44	.13	3	51	15	4-8	
	222	331	5	A	.94	1.14	.13	3	51	15	3-7	
	230	332	5	C	.57	-.85	.13	3	51	15	4-8	
SCAT I 2B	105	333	1	A	1.93	1.25	.11	1	52	11	2-6	12
	89	334	1	A	1.12	.64	.20	1	52	11	2-6	
	81	335	1	D	1.72	.51	.19	1	52	11	2-6	
	71	336	1	B	1.90	.18	.21	1	52	11	2-6	
	9	337	1	B	.74	1.72	.15	1	52	11	2-6	
	7	338	1	E	.56	1.80	.15	1	52	11	2-6	
	354	339	5	D	1.41	1.36	.12	3	52	11	5-9	
	289	340	5	E	1.93	.41	.10	3	52	11	4-8	
	292	341	5	C	1.36	.43	05	3	52	11	3-7	
	219	342	5	C	.73	1.20	.15	3	52	11	4-8	
	216	343	5	A	.76	1.26	.15	3	52	11	3-7	
SCAT I 3A	110	344	1	D	1.88	1.47	.19	1	53	22	2-6	13
	97	345	1	E	2.10	.83	.05	1	53	22	2-6	
	57	346	1	E	.95	-.03	.14	1	53	22	2-6	
	52	347	1	E	.82	-.15	.14	1	53	22	2-6	
	65	348	1	C	2.10	.05	.21	1	53	22	2-6	
	34	349	1	B	.00	-.61	.15	1	53	22	2-6	
	15	350	1	C	.69	1.09	.14	1	53	22	2-6	

Item Source	Index #	Rec. #	Instr. Set	Ans.	IPA	IPB	IPC	Type	Form Code	# Items Cont.	Line	Index
PSAT OPT	704	109	11	C	.79	1.63	.1	4	68	12	2-6	
	713	110		B	1.92	1.88	.23					
SAT PSA 43												
	296	86	7	C	1.39	.49	.14	3	69	13	5-9	P43
	373	87		A	1.74	1.77	.14				3-7	
	387	88		B	1.03	2.37	.08				4-8	
	120	89	3	A	1.27	.15	.14	2			2-6	
	597	90	11	D	.85	.20	.14	4				
	746	91		A	.65	2.79	.08					
	368	92	7	B	.84	1.59	.14	3			5-9	
	381	93		C	1.19	2.12	.09				3-7	
	152	94	3	B	1.55	1.48	.18	2			2-6	
	142	95		E	1.76	1.13,	.19					
	170	96		A	1.27	2.24	.14					
	677	97	11	B	1.23	1.19	.15	4				
	727	98	11	C	1.27	2.21	.18					
SAT QSA43												
	378	72	7	E	3.00	1.94	.10	3	70	14	5-9	Q43
	376	73		E	1.31	1.89	.08				4-8	
	134	74	3	E	2.55	.77	.15	2			2-6	
	128	75	3	C	.97	.59	.14	2				
	179	76		B	2.32	2.44	.17					
	702	77	11	A	.91	1.61	.14	4				
	319	78	7	A	1.10	.89	.14	3			3-7	
	371	79		C	2.05	1.74	.17	3			5-9	
	384	80		A	1.55	2.28	.12				4-8	
	157	81	3	C	2.14	1.83	.21	2			2-6	
	185	82		E	1.35	2.88	.19					
	662	83	11	D	.95	.89	.14	4				
	710	84		D	1.31	1.76	.14					
	717	85		A	1.03	1.93						
GRE QCR1	722	3	11	A	.83	2.05	.17	4	71	30	2-6	G1
	752	4		C	.58	3.20	.14					
	739	5		B	.66	2.44	.17					
	160	6	3	B	1.02	1.97	.21	2				
	188	7		A	1.55	3.30	.17					

GRE QGR1	186	8	3	A	1.05	3.05	.14	2	71	30	2-6	G1
	202	9		A	.80	4.76	.17					
	405	10	7	C	1.31	3.67	.23	3			5-9	
	401	11		D	.88	3.34	.19				3-7	
	394	12		A	.59	2.82	.14				4-8	
	398	13		C	.77	3.12	.14				5-9	
	393	14		B	1.06	2.80	.19				3-7	
	563	15	11	E	.53	- .12	.17	4			2-6	
	719	16		B	.57	1.97	.17					
	753	17		D	.55	3.35	.17					
	750	18		B	.53	3.06	.17					
	756	19		D	.74	3.82	.19					
	758	20		E	.59	3.87	.17					
	148	21	3	B	.84	1.37	.17	2				
	199	22	3	D	.95	4.21	.20	2				
	192	23		E	1.65	3.59	.19					
	197	24		E	1.51	3.87	.09					
	200	25		A	.97	4.66	.15					
	198	26		B	.90	4.12	.20					
	335	27	7	E	.84	1.11	.15	3			6-10	
	389	28		D	.64	2.66	.17				4-8	
	407	29		E	.74	3.86	.07					
	396	30		D	.84	2.91	.17				5-9	
	382	31		A	.82	2.17	.17				3-7	
	392	32		B	.55	2.78	.17				5-9	
GRE KQGR 2												
	744	33	11	A	.63	2.58	.16	4	72	39	2-6	G2
	742	34		C	.78	2.55	.03					
	735	35		E	.69	2.37	.16					
	749	36		E	1.13	3.04	.27					
	754	37		A	1.02	3.44	.22					
	182	38	3	C	.59	2.61	.19	2				
	163	39		B	1.40	2.08	.27	2				
	187	40		E	.77	3.16	.26					
	184	41		C	1.01	2.86	.19					
	183	42		A	.59	2.80	.16					
	191	43		E	.57	3.49	.19					

INSTRUCTION SETS

INSTRUCTION SET NO.	TEXT
1 with 5 lines	This question has one word followed by five words or phrases lettered A,B,C,D and E. Read the word. Then pick the lettered word or phrase that has the same or almost the same meaning.
2 with 2 lines	Pick the word or phrase that has the same or almost the same meaning as the first word.
3 with 4 lines	This question has one word followed by 5 words or phrases lettered A through E. Read the word. Then pick the lettered word or phrase that is most nearly opposite in meaning.
4 with 2 lines	Pick the word or phrase that is most nearly opposite in meaning to the first word.
5 with 6 lines	This question has a sentence in which one word is missing; a blank space indicates where the word has been removed from the sentence. Beneath the sentence are five words lettered A, B, C, D, and E one of which is the missing word. You are to select the missing word by deciding which one of the five words best fits in with the meaning of the sentence.
6 with 2 lines	Select the missing word which best fits in with the meaning of the sentence.
7 with 5 lines	This sentence has one or more blank spaces, each blank indicating that a word has been omitted. Beneath the sentence are 5 lettered words or sets of words. You are to choose the one word or set of words which when inserted in the sentence, best fits in with the meaning of the sentence as a whole.

INSTRUCTION SET NO.	TEXT
8 with 2 lines	Select the word or set of words which best completes the following sentence.
9 with 6 lines	This question begins with 2 words. These two words go together in a certain way. Under them there are 4 other pairs of words lettered A, B, C, D. Find the lettered pair of words that go together in the same way as the first pair of words.
10 with 3 lines	Find the lettered pair of words that go together in the same way as the first pair of words.
11 with 5 lines	In this question a related pair of words or phrases is followed by 5 lettered pairs of words or phrases. Select the lettered pair which best expresses a relationship similar to that expressed in the original pair.
12 with 3 lines	Select the lettered pair of words or phrases which best expresses a relationship similar to that expressed in the original pair.
13 with 5 lines	In this question, the first sentence is followed by an incomplete statement and 4 suggested answers. lettered A, B, C, and D. You are to decide which one of these answers is best. Your choice should be based on what the first sentence says.
14 with 5 lines	In this question, the first sentence is followed by an incomplete statement and 4 suggested answers lettered A, B, C, D. You are to decide which one of these answers is best. Your choice should be based on what the first sentence says.

Appendix C

Item Selection Tables

Following are the item selection tables employed in Forms A and B of the BRTT. Together with the item data reported in Appendix B they constitute a description of the test as used in the current study.

```
ENTER FILENAME . BPTTAB TBL

5
  0   0   0   3   3  12  12  63  63  87  87 108 108 172 172   0   0   0   0   0
  0   0   0 604 604 615 613 670 670 667 667 371 371   0   0   0   0   0   0   0
  0   0   0 429 429 512 512 566 566 641 641 712 712 741 741   0   0   0   0   0
  0   0   0 782 782 818 818 868 868 891 891 379 379 384 384   0   0   0   0   0
  0   0   0 428 428 496 496 539 539 677 677 690 690 740 740   0   0   0   0   0
  0   0   0   4   4  11  11  77  77 105 105 154 154 179 179   0   0   0   0   0
  0   0   0 601 601 645 645 673 673 651 651 372 372   0   0   0   0   0   0   0
  0 419 419 447 447 464 464 579 579 635 635 706 706   0   0   0   0   0   0   0
  0   0   0 734 734 828 828 362 362 347 347 334 334 381 381   0   0   0   0   0
  0 423 423 428 428 474 484 502 502 617 617 656 656   0   0   0   0   0   0   0
  0   0   0   2   2  13  13  47  47  81  81 132 132 184 184 191 191 157 157   0
  . 778 778 779 779 844 844 289 289 304 304 364 364 385 385   0   0   0   0   0
  0   0   0 433 433 468 464 584 584 663 663 688 688 729 729   0   0   0   0   0
  0 771 771 784 784 812 812 239 279 371 371 378 378 395 395 408 408   0   0   0
      0   0 443 443 500 500 567 567 613 613 704 704 727 727 746 746   0   0   0
  0   0   0   7   7  15  15  37  37 124 124 160 160 187 187 188 188 200 200   0
  0 772 772 750 750 212 212 284 284 291 291 373 373 396 396 407 405   0   0   0
411 411 411 427 427 519 519 602 602 633 633 720 720 745 745   0   0   0   0   0
  0 765 765 751 751 228 228 248 248 342 342 360 360 391 391 401 401   0   0   0
421 424 424 451 451 527 527 552 552 667 667 696 696 725 725 731 751 759 759   0
  0       1   3   3  18  18  77  97 141 141 148 148 181 181 188 188 195 195 207
772 772 772 207 207 218 218 254 254 327 327 356 356 386 386 395 395 405 405   0
  0 418 418 433 433 524 524 542 542 673 673 714 714 727 727 732 732 737 737   0
741 747 753 203 203 224 224 289 289 308 308 362 362 387 381 398 398 402 402   0
  0 413 413 472 472 480 480 604 604 637 637 719 719 744 744 777 771 778 778   0
  0   0   0   1   1   1   1   1   1   1   1   1   1   1   1   0   0   0   0   0
  0   0   0   2   1   3   3   1   1   1   3   1   1   0   0   0   0   0   0   0
  0   0   0   4   4   4   4   4   4   4   4   4   4   4   4   0   0   0   0   0
  0   0   0   3   3   3   3   3   3   3   5   3   3   3   3   0   0   0   0   0
  0   0   0   4   4   4   4   4   4   4   4   4   4   4   4   0   0   0   0   0
      0   0   1   1   1   1   1   1           2   2   2   2   0   0   0   0   0
```

```
0    4    4    4    4    4    4    4    4    4    4    4    4    0    0    0    0    0    0    0

0    0    0    3    3    3    3    3    3    3    3    3    3    3    3    0    0    0    0    0

0    4    4    4    4    4    4    4    4    4    4    4    4    0    0    0    0    0    0    0

0    0    0    1    1    1    1    1    1    1    1    1    2    2    2    2    2    2    2    0

0    3    3    5    3    5    5    3    3    3    3    3    3    3    3    0    0    0    0    0

0    0    0    4    4    4    4    4    4    4    4    4    4    4    4    0    0    0    0    0

0    3    3    3    5    5    5    3    3    3    3    3    3    3    3    3    3    0    0    0

0    0    0    4    4    4    4    4    4    4    4    4    4    4    4    4    4    0    0    0

0    0    0    1    1    1    1    1    1    2    2    2    2    2    2    2    2    2    2    0

0    5    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    0    0    0

4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    0    0    0    0    0

0    5    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    0    0    0

4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    0

0    1    1    0    0    1    1    1    1    2    2    2    2    2    2    2    2    2    2    2

3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    0

0    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    0

3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    0

0    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    4    0

STOP --

READY
```

BRTT2A (Continued)

```
ERT

N TABLE

UTPUT DEVICE  (6-PRINTER 7-[illegible]   7

NTER FILENAME   ERT/26.TBL

 0   0   0   1   1  26  26  41  41 101 101 110 110   0   0   0   0   0   0   0
 0   0   0 506 506 612 714 574 574 550 550 597 597   0   0   0   0   0   0   0
 0   0   0 445 445 451 451 573 573 664 664 651 651   0   0   0   0   0   0   0
 0   0   0 777 777 611 611 676 676 675 575 576 576   0   1   0   0   0   0   0
 0   0   0 457 457 511 511 611 611 677 672 711 711   0   0   0   0   0   0   0
 0   0   0   6   6  17  12  71  71  85  85 110 110 185 185   0   0   0   0   0
 0 767 767 757 757 624 624 660 660 256 256 277 277 266 266   0   0   0   0   0
 0   0   0 424 424 520 520 545 545 626 626 710 710 721 721   0   0   0   0   0
 0 779 779 756 756 540 540 671 671 215 215 274 274  18 780   0   0   0   0   0
 0   0   0 420 420 517 517 575 575 674 674 700 700 747 747   0   0   0   0   0
 0   0   0   5   5  14  14  55  75 140 143 157 157 170 170 151 151   0   0   0
 0   0   0 721 721 605 605 651 651 216 216 271 271 351 351   0   0   0   0   0
27 427 427 421 421 504 504 551 551 662 667 716 716 742 742   0   0   0   0   0
 0 774 774 775 775 212 212 275 275 211 211 266 266 257 257 757 757   0   0   0
.4 421 421 455 455 476 476 546 546 651 651 717 717 741 741   0   0   0   0   0
 0   0   0  10  10  15  16  51  51 126 126 161 161 176 176 167 167 156 156   0
 0 771 771 762 762 226 226 256 256 300 300 376 376 381 381 403 403   0   0   0
 0 421 421 441 441 455 455 557 557 650 650 654 654 711 711 745 745   0   0   0
 0 765 765 240 240 222 222 241 241 225 225 366 366 750 750 404 404   0   0   0
 0 417 417 421 421 471 471 560 560 664 664 791 791 716 716 755 755 760 760   0
 0   0   0   6   6  34  34 120 120 125 175 170 170 183 183 151 151 201 201   0
 0 755 764 744 735 ; 5 415 434 754 331 382 360 360 383 383 400 400   0   0   0
14 426 426 417 415 524 524 742 682 674 677 717 717 729 715 774 774 776 776   0
 0 744 764 734 736 220 220 274 274 225 225 232 232 384 384 401 401 407 407   0
11 417 416 441 441 502 502 425 425 646 646 722 722 746 745 734 779 761 761   0
 0   0   0   1   1   .   1   .   1   .   1   .   1   0   0   0   0   0   0   0
     0   5   5   5   7   5   5   7   5   5   7   0   0   0   0   0   0   0   0
 0   0   0   -   2   4   4   4   4   4   4   4   4   0   0   0   0   0   0   0
```

```
.   .   .   .           .   .   .   .   .   .   .   .   .   .   .   .   .
0   0   0   4   4   4   4   4   4   4   4   4   4   0   0   0   0   0   0
0   0   0   1   1   1   1   1   1   1   1   1   1   2   2   0   0   0   0
0   5   5   5   5   5   3   5   5   3   3   3   3   3   3   0   0   0   0
0   0   0   4   4   4   4   4   4   4   4   4   4   4   4   0   0   0   0
0   5   5   5   5   5   5   5   3   3   3   3   3   3   3   0   0   0   0
0   0   0   4   4   4   4   4   4   4   4   4   4   4   4   0   0   0   0
0   0   0   1   1   1   1   1   1   2   2   2   2   2   2   2   2   0   0
0   0   0   5   5   5   5   5   5   3   3   3   3   3   3   0   0   0   0
4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   0   0   0   0
0   5   5   5   5   5   5   3   3   3   3   3   3   3   3   3   3   0   0
4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   0   0   0   0
0   0   0   1   1   1   1   1   1   2   2   2   2   2   2   2   2   2   2
0   5   5   5   5   3   3   3   3   3   3   3   3   3   3   3   3   0   0
0   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   0   0
0   5   5   3   3   3   3   3   3   3   3   3   3   3   3   3   3   0   0
0   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
0   0   0   1   1   1   1   2   2   2   2   2   2   2   2   2   2   2   2
0   5   5   5   5   3   3   3   3   3   3   3   3   3   3   3   3   0   0
4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
0   5   5   5   5   3   3   3   3   3   3   3   3   3   3   3   3   3   3
4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
STOP --
READY
```

BRTT2B (Continued)

REFERENCES

Ahl, D. H. Survey of public attitudes toward computers in society. Creative Computing, Nov.-Dec. 1975, 49-51.

Angoff, W. H., & Huddleston, E. M. The multi-level experiment. A study of a two-stage test system for the College Board Scholastic Aptitude Test. Statistical Report 58-21. Princeton, N. J.: Educational Testing Service, 1958.

Alderman, D. Evaluation of the TICCIT computer assisted instructional system in the community college (PR-78-10). Princeton, N. J.: Educational Testing Service, September 1978.

Atkinson, J. W. (Ed.) Motives in fantasy, action, and society. Princeton, N. J.: Van Nostrand, 1958.

Atkinson, J. W., & Litwin, G. H. Achievement motive and test anxiety conceived as motive to approach success and motive to avoid failure. Journal of Abnormal and Social Psychology, 60, 52-63.

Bayroff, A. G., & Seeley, L. C. An exploratory study of branching tests. Technical Research Note 188. U.S. Army Behavioral Science Research Laboratory, June 1967.

Betz, N. E., & Weiss, D. J. An empirical study of computer-administered two-stage abiity testing. Research Report 73-4. Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota, 1973.

Betz, N. E., & Weiss, D. J. Simulation studies of two-stage ability testing. Research Report 74-4. Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota, 1974.

Betz, N. E., & Weiss, D. J. Effects of immediate knowledge of results and adaptive testing on ability performance. Research Report 76-3. Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota, June 1976. (NTIS No. ADA027147) (a)

Betz, N. E., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive testing. Research Report 76-4. Minneapolis: Department of Psychology, Psychometric ethods Program, University of Minnesota, 1976. (NTIS No. ADA027170) (b)

Birnbaum, A. Some latent trait methods and their use in inferring an examinee's ability. In Lord and Novick Statistical theories of mental test scores. New York: Addison-Wesley Publishing Co., 1968.

Cartwright, G. F., & Derevensky, J. L. An attitudinal study of computer-assisted testing as a learning method. Psychology in the Schools, July 1976, 13(3), 317-321.

Cleary, T. A., Linn, R. L., & Rock, D. A. An exploratory study of programmed tests. Educational Psychological Measurement, 1968, 28, 345-360.

De Finnetti, B. Methods of discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 1965, 18, 87-123.

Gaudry, E., & Spielberger, D. C. Anxiety and educational achievement. Sydney: John Wiley and Sons, 1971.

Green, B. F., Jr. Discussion. Proceedings of the First Conference on Computerized Adaptive Testing. Professional Series 75-6. Personnel Research and Development Center, U. S. Civil Service Commission, Washington, D.C., 1975.

Gulliksen, H. Theory of Mental Tests. New York: Wiley, 1950.

Hájek, J. Nonparametric Statistics. San Francisco: Holden-Day, 1969.

Hansen, J. B. Effects of feedback, learner control, and cognitive abilities on state anxiety and performance in a computer-assisted instruction task. Journal of Educational Psychology, 1974, 66(2), 247-254.

Harrison, R. Thematic Apperception Methods. In B. B. Wolman (Ed.) Handbook of Clinical Psychology. New York: McGraw-Hill, 1975. Pp. 562-620.

Hedl, J. J., and others. Affective reactions toward computer-based intelligence testing. Journal of Consulting and Clinical Psychology, April 1973, 40(2), 217-222.

Hick. W. E. Information theory and intelligence tests. British Journal of Psychological Statistics, 1951, Sect. 4, 157-164.

Hutt, M. L. A clinical study of "conservative" and "adaptive" testing with the revised Stanford-Binet. Journal of Consulting Psychology, 1947, 11, 93-103.

Johnson, D. F., & Mihal, W. L. Performance of blacks and whites in computerized versus manual testing environments. American Psychologist, August 1973, 694-699.

Jones, D. J. Personal communication, 1978.

Koch, B., & Patience, W. Student attitudes toward tailored testing. In D. J. Weiss (Ed.) Proceedings of the 1977 Computerized Adaptive Testing Conference. (N00014-76-C-0243) Arlington, Va.: Office of Naval Research, 1978.

Kreitzberg, C. B., Stocking, M. L., & Swanson, L. *Computerized adaptive testing: Principles and directions*. Princeton, N. J.: Educational Testing Service, February 1978.

Larkin, D. C., & Weiss, D. J. *An empirical investigation of computer-administered adaptive ability testing. Research Report 74-3*. Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota, 1974.

Levitt, E. E. *The psychologically of anxiety*. Indianapolis: Bobbs-Merrill, 1967.

Linn, R. L., Rock, D. A., & Cleary, T. A. Sequential testing for dichotomous decision. *Educational Psychological Measurement*, 1972, *32*, 85-96.

Locke, E. A., Carledge, N., & Koeppel, J. Motivational effects of knowledge of results: A goal-setting phenomenon? *Psychological Bulletin*, 1968, *70*(6), 474-485.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.) *Computer Assisted Instruction, Testing and Guidance*, Chapter 8. New York: Harper & Row, 1970.

Lord, F. M. A Broad-Range Tailored Test of Verbal Ability. *Applied Psychological Measurement*, 1977, *1*, 95-100. (a)

Lord, F. M. *Some how and which for practical tailored testing*. (Pre-publication draft) Princeton, N. J.: Educational Testing Service, 1977. (b)

Lord, F. M. *Applications of item response theory to practical problems*. Hillsdale, N. J.: Erlbaum Publishers, 1980 (in press).

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. New York: Addison Wesley, 1968.

Mandler, G., & Sarason, S. B. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, *47*, 166-173.

McBride, J. R. *Research on adaptive testing, 1973-1976: A review of the literature*. University of Minnesota, 1976.

Means, R. S., & Means, G. H. Achievement as a function of the presence of prior information concerning aptitude. *Journal of Educational Psychology*, 1971, *63*(3) 185-87.

Novick, M. R. *Bayesian methods in psychological testing*. Research Bulletin RB-69-31. Princeton, N. J.: Educational Testing Service, 1969.

Noyce, R. N. Microelectronics. Scientific American, 1977, 237(3), 63-69.

Olivier, P. An evaluation of the self scoring f.exilevel testing model. Unpublished doctoral dissertation, Florida State University, 1974.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Pine, S. Reduction of test bias by adaptive testing. In D. J. Weiss (Ed.) Proceedings of the 1977 Computerized Adaptive Testing Conference. (N00014-76-C-0243). Arlington, Va.: Office of Naval Research, 1978.

Prestwood, J. Effects of knowledge of results and varying proportion correct on ability test performance and psychological variables. In D. J. Weiss (Ed.), 116-127, Proceedings of the 1977 Computerized Adaptive Testing Cference. (N00014-76-C-0243). Arlington, Va.: Office of Naval Research, 1978.

Rosenthal, R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1966.

Sarason, I. G. Experimental approaches to test anxiety: Attention and the uses of information. In C. D. Spielberger, 1972, op cit.

Schmidt, F., Urry, V., & Gugel, J. Computerized tailored testing: Examinee reactions and evaluations. Washington, D.C.: U.S. Civil Service Commission PB-276 748, December 1977.

Spence, K. W. Anxiety (drive) level and performance in eyelid conditioning. Psychological Bulletin, 1964, 61, 129-139.

Strang, H. R., & Rust, J. O. The effects of immediate knowledge of results and task definition on multiple-choice answering. Journal of Experimental Education, 1973, 42(1), 77-80.

Swanson, L., Stocking, M. Computerized adaptive testing facility: System. Princeton, N. J.: Educational Testing Service, December 1977.

Sympson, J. A model for testing with multidimensional items. In D.J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference, (N00014-76-C-0243). 82-9B. Office of Naval Reserach, 1978.

Urry, V. W. Computer-assisted testing with live examinees: A rendevous with reality. Personnel Research and Development Center, U.S. Civil Service Commission, 1976.

Urry, V. W. *A multivariate sampling procedure and a method of multidimensional tailored testing*. Paper presented at the 1977 Computerized Adaptive Testing Conference, University of Minnesota, 1977.

Urry, V. W. Five years of research: Is computer-assisted testing feasible? *Proceedings of the First Conference on Computerized Adaptive Testing*. Professional Series 7556. Personnel Research and Development Center, U.S. Civil Service Commission, 1975.

Urry, V. Paper presented at the Fifth Annual Conference Educational Assessment of the International Association for Educational Assessment, Educational Testing Service, Princeton, N. J., May 27-31, 1979.

Vale, C. D., & Weiss, D. J. *A Study of computer-administered stradaptive ability testing. Research Report 75-4*. Minneapolis, Minnesota: Department of Psychology, Psychometric Methods Program, University of Minnesota, 1975.

Walker, R. E., Neilsen, M. Kay, & Nicolay, R. C. The effects of failure and anxiety on intelligence test performance. *Journal of Clinical Psychology*, 1965, *20*, 400-402.

Waters, B. *Empirical investigation of the stradaptive testing model for the measurement of human ability*. Ph.D. Dissertation, Florida State University, 1974.

Weiner, A. S., & Adams, W. V. The effect of failure and frustration on reflective and impulsive children. *Journal of Experimental Child Psychology*, 1974, *17*, 353-359.

Weiner, G. The effect of distrust on some aspects of intelligence test behavior. *Journal of Consulting Psychology*, 1957, *21*(2), 127-120.

Weiss, D. J., & Betz, N. E. *Ability Measurement: Conventional or adaptive? Research Report 73-1*. Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota, 1973.

Weiss, D. J. *The stratified adaptive computerized ability test. Research Report 73-3*. Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota, 1973.

Weiss, D. J. *Strategies of adaptive measurement. Research Report 74-5*. Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota, 1974.

Weiss, D. J. Adaptive testing research at Minnesota: Overview, recent results and future directions. Proceedings of the First Conference on Computerized Adaptive Testing. Professional Series 75-6. Personnel Research and Development Center, U.S. Civil Service Commission, Washington, D.C., 1975.

Weiss, D. J. Remarks. Proceedings of the 1977 Computerized Adaptive Testing Conference, University of Minnesota, 1977, 442-443.

Wickes, T. A., Jr. Examiner influence in a testing situation. Journal of Consulting Psychology, 1956, 20, 23-26.

Wine, J. D. Test anxiety and direction of attention. Psychological Bulletin, 1971, 76(2), 92-104.

Wood, R. Response-Contingent Test. Review of Educational Research, 1973, 43, 529-544.