AN EXAMINATION OF THE PRECISION OF MEASUREMENT OF COMPUTERIZED ADAPTIVE TESTS WITH LIMITED ITEM POOLS

by

PERRY N. HALKITIS

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1995

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

4/12/95
Date

David Rindskopf
Chair of Examining Committee
David M. Rindskopf

4/12/95
Date

Executive Officer
Alan L. Gross

Professor David Rindskopf

Professor Alan L. Gross

Professor Roger E. Millsap

Supervisory Committee

The City of University of New York

Abstract

AN EXAMINATION OF THE PRECISION OF MEASUREMENT OF COMPUTERIZED ADAPTIVE TESTS WITH LIMITED ITEM POOLS

by

Perry N. Halkitis

Adviser: David M. Rindskopf, Ph.D.

This paper describes a method for examining the precision of a computerized adaptive test. When item pools are ideal in composition, CAT has been proven to increase efficiency and accuracy in both live and simulated testing situations. The precision of measurement of a CAT, based on IRT methodologies, is a function of the information that is provided by the individual items and the overall item pool. When pools are limited in their composition the results are less ideal. The precision and efficiency of the CAT process may be limited by the test construction process and speaks to the limitations of CAT in environments where the generation of large and extensive item pools is not possible. To further examine the precision of a CAT, standard errors of measurement ascertained in the testing of simulees in a CAT with a limited item pool were compared to those results obtained in a live paper-and-pencil achievement testing of 4371 nursing students on four versions of an examination of calculations of drug administration. Item difficulties and examinee abilities were ascertained with maximum likelihood estimation procedures based on the Rasch model. Analyses examined the extent to which indices of precision were enhanced when this achievement testing was administered in a CAT format given the realities of an item

pool. CAT measures of precision were considered when the simulated examinee pools were uniform and normal. Results suggest that regardless of the size of the item pool, CAT provides greater precision in measurement with a smaller number of items administered even when the choice of items is limited, but fails to achieve equiprecision along the entire ability continuum.

ACKNOWLEDGMENTS

Success in a doctoral program depends on the support and enthusiasm of your mentors. For their belief in me and in their work and for their humanity, I extend thanks to Drs. Alan Gross, David Rindskopf, and Roger Millsap. They have been my inspirations and my guides throughout this demanding and rewarding journey. Much gratitude to Drs. Howard Everson and Louis Primavera for serving as my outside readers. For sharing their insights and knowledge for the last several years, thanks to Drs. Carol Tittle and Zita Cantwell.

I wish to acknowledge my colleagues at the Hunter College Campus Schools, National League for Nursing, and Professional Examination Service for their calm understanding.

Many thanks to Amy Schmidt, my buddy in the doctoral program, who provided the drive to push even harder at every step and provided a summer of fun preparing for comps. For helping me through my toughest moments and allowing me see the wisdom of psychodynamic theories, much appreciation to Dr. Nancy Eisenman. To my friends--Mala, Marc, Eileen, Arthur, Frank, Jim, George, Anne, Kathleen, Pat, Cliff, Jonathan, Rick, Greg, Tamar, Carol, Harris--thanks for the friendship and honesty and especially the laughs. And to my many pals who have lost their lives to this ridiculous thing called AIDS, a kiss and a prayer. (Maybe one day, before it's too late, society will wake up and realize what is happening!)

Of course to my mom, Kalliope Halkitis, endless respect and love for among other things just being there for me ALWAYS. And for rounding out my family, love to my little brother Tony, my sister-in-law Lisa, and to my empathetic and talented cousin Ellaina.

To the future generation, my girls Sophia, Lucy, and Sadie, thanks for reminding me about the important things in life.

To my father, Nicholas Halkitis, and my doppelganger, Robert Massa, whom life swept away much too early and whom I miss more and more every day, I owe the world. Without their love and their passion and their pride, none of this would have any meaning. This dissertation is dedicated in everlasting loving memory to them.

In the end, the completion of this project was difficult. For his inspiration to draw the project to a close, for making this document look great, and for filling in the final and most important piece of life's puzzle, eternal gratitude and love to my life partner Jeff Bosacki. I'm ready; let's go!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

INTRODUCTION

For approximately one century, paper-and-pencil, fixed length examinations have been a mainstay in educational settings. Based on the ideas first documented by E. L. Thorndike (1904), these examinations have provided an opportunity to test large numbers of people efficiently. The popularity of paper-and-pencil examinations throughout this century can be attributed to the relatively weak assumptions of classical test theory, the psychometric model upon which these examinations are based, and the lack of sophisticated technology to handle complex calculations (Hambleton & Jones, 1993). Yet two developments, one technological and the other psychometric, have provided the avenue for reconsidering the methods utilized to test individuals (Crocker & Algina, 1986). The availability of large computers has enabled psychometricians to implement item response theory (IRT). This has led to increasingly sophisticated research aimed at improving testing procedures (Weiss & Kingsbury, 1984).

Adaptive testing, the theoretical foundation for computerized adaptive testing (CAT), incorporates procedures by which items are tailored to an individual based on his or her responses to previous questions. Many examinations, including intelligence examinations have incorporated such methodologies for years, whereby items are chosen in light of earlier performance. More often, the later questions are ones the examinee's previous responses indicate will be neither too easy nor too difficult for the examinee (Cronbach, 1990). In a full-fledged implementation of the process, the examinee's ability is estimated , based on all the questions the examinee has answered

thus far (Ward, 1985).

Computerized adaptive testing (CAT), a practical application of IRT, uses similar procedures yet implements the element of the computer as a medium for presentation and item selection. The adaptive nature of CAT is also referred to in the literature as branched testing, individualized testing, programmed testing, sequential item testing, response-contingent testing, and computerized testing (Lord, 1980; Weiss, 1985). But CAT is more than just adaptive testing. It is rather an orientation to the testing process that brings together the technology and the "newer" psychometric framework of IRT with the theoretical impetus of "individualized" testing in what is suggested to be a more efficient and accurate approach to the testing process.

The idea of CAT is not a new one. The addition of the computerized element to adaptive testing was first suggested by William W. Turnbull in 1951 (Lord, 1980). Since then, the idea of computerized adaptive testing was explored in theory for years (Angoff & Huddleston, 1958; Cleary, Linn, & Rock, 1968; Lord, 1970; Wright, 1968), and has gained increased momentum in recent years. In 1985, William Ward suggested that the future of CAT was already upon us. And less than one decade later, the future known as computerized adaptive testing is a reality for many including nursing students who prepare to become licensed in the United States (Halkitis & Leahy, 1993).

It has been argued that CAT provides a vehicle for measurement that is superior to conventional fixed-length tests simply because it is more efficient and more precise in the determination of the abilities of examinees. The argument of precision is based on the notion that when a test is administered in a computerized adaptive format, individuals are presented items that maximize information at their own ability levels.

The result is an individualized or tailored test for each examinee with maximum information, and as a result greater precision in measurement. It is claimed that fewer than half as many questions are needed as in conventional testing, and a CAT yields broad range accuracy in assessing the ability of examinees (Ward, 1985).

The procedures utilized in CAT are similar to those employed when individual intelligence tests, such as the Stanford-Binet, are administered. The main difference rests on the speed, precision, and computer administration involved in CAT (Anastasi, 1985).

A number of studies, either directly or indirectly, have examined the possible effects of CAT on the precision of measurement in order to establish the effectiveness of this approach. Many of these studies have either incorporated theoretical, infinite item pools or items pools that are finite but ideal in their composition. Still others have utilized item pools based on data from actual examinations that are modified in some respect in order to adhere to the purposes of the research. Few have used actual unmodified data in assessing the issue of precision.

This investigation examined the precision of measurement of a computerized adaptive examination when an item pool is limited by the realities of the test construction process. Is the standard error of ability estimate ascertained in the administration of a CAT equal to or less than that of a conventional test given the confines of a real item pool? To what extent does the composition of the item pool utilized in a CAT affect the precision with which the examination measures ability? It was hypothesized that the limitations of a real item pool will result in less dramatic advantages than have been shown to be theoretically possible.

In the end, an examination of CAT procedures using an existing item pool

limited by the realities of the test construction process provides a check regarding the claims of precision being put forth regarding computerized adaptive testing, and speaks to the realities of this testing framework, where unlike in large testing programs, ideal situations are not possible.

In the sections that follow, overviews of Item Response Theory and adaptive testing will provide the basis for considering the workings of CAT. Following a discussion of the principles involved therein is an examination of the relevant research in terms of the issue of precision of measurement. Finally, the methodology, results, and conclusions are presented.

Chapter II

REVIEW OF LITERATURE

A. The Psychometric Framework of CAT: Item Response Theory

Item response theory (IRT), latent trait theory, or item characteristic curve theory, has provided the psychometric impetus for the development of computerized adaptive testing. While the set of models postulated by this psychometric framework can be traced back to as early as the 1940's (Tucker, 1946), the application and development of this theoretical orientation in measurement can be attributed to the pioneering work of Frederick Lord (1952; 1953). It was not until the late 1960's and 1970's that the application of IRT models to educational measurement became apparent. The works of Lord and Novick (1968), Wright (1968), Panchapakesan (1969) and Samejima (1972) helped introduce this framework to educational testing.

Item response theory is a family of mathematical descriptions about what happens when an examinee meets an item (Thissen & Mislevy, 1990), and suggests that in testing situations the performance of any examinee can be predicted by defining the examinee's traits or abilities (Harris, 1989; Lord & Novick, 1968). An IRT model specifies a relationship between the examinee test performance and the latent trait that the examinee possesses.

Unlike classical test theory, IRT models allow us to predict how an individual will respond to items that appear on an examination, given that the traits that underlie the performance are fewer in number than the items that constitute the test. And, in fact, most IRT models assume unidimensionality, suggesting that one trait or ability

underlies performance on any particular examination.

The central concept of IRT is the item characteristic curve (ICC), a function that represents the probability of answering a given item correctly given an examinee's latent trait or ability. It is the non-linear regression function of item score on the latent trait being measured by the test (Hambleton & Cook, 1977). The function, often depicted as an *S*-shaped curve, is a mathematical statement as to how response depends on the individual's level of ability or skill (Lord, 1980). The particular mathematical model which is postulated is dependent on the assumptions the investigator is willing to make about the test data; different models are specified depending on the assumptions that an individual is willing to make. The only consideration in selecting a latent trait model is whether or not the data that are to be used in the analysis satisfy the assumptions of the model (Hambleton & Cook, 1977).

Three mathematical functions are most commonly used to represent the ICCs. While all three functions possess common characteristics, what varies is the number of parameters that they use to specify the ICC. The one, two, and three parameter logistic models respectively utilize one, two, and three item characteristics to describe the ICC by combining these parameters into a logistic function to relate examinee ability and the item parameter(s) to the probability of correctly responding to an item (Harris, 1989).

Item difficulty, item discrimination, and the pseudo-guessing parameter are the characteristics of items considered in IRT models. For the 1-PL and 2-PL models, the item difficulty (b), the point of inflection on the ability scale, represents the level along the ability continuum at which an item has a 50% probability of a correct response. It can be thought of additionally as the point at which a person with the ability level

equal to the difficulty of the item has a 50% chance of answering the item correctly, or the point at which 50% of the examinees at the ability level will answer the item correctly. For the 3-PL model, item difficulty is defined as the point at which the probability of correctly answering an item is equal to $(1 + c)/2$, where c represents the lower asymptote of the ICC. The item discrimination (a) reflects the slope of the ICC at the difficulty level of the item. The pseudo-guessing parameter (c) is the lower asymptote of the ICC and reflects the probability of a correct response by examinees of an infinitely low ability level. Typically, the pseudo-guessing (or pseudo-chance) parameter assumes values that are smaller than the value that would result if examinees of low ability were to randomly guess a response to the item. Lord (1974) suggests that this phenomenon may be due to the ingenuity of item writers to write "attractive" distractors.

In the one-parameter logistic (1-PL) or Rasch model (Rasch 1960; Wright, 1968; Wright, 1977), item characteristic curves all possess the same steepness or slope, thus suggesting that all items on an examination have the same discriminating power. In addition, guessing is not incorporated into the model, and thus all curves are suggested to have a lower asymptote approaching 0.0. What varies from item to item is the difficulty as indicated by the parameter b. More difficult items are associated with a larger b parameter; easier ones with a smaller b value. The equation for the ICC of the 1-PL or Rasch model is

$$P_i = P_i(\Theta) = e^{Da'(\Theta - bi)}/(1 + e^{Da'(\Theta - bi)})$$

where $P(\Theta)$ is the probability of an examinee with ability Θ answering item *i* correctly; D represents a scaling constant to maximize correspondence between the normal and logistic ogives (usually set at 1.7); a′ represents the common discriminating power of

all items, and b_i represents the difficulty of item *i*. In the 1-PL model, item difficulty is defined as the value at which the probability of a correct response is equal to 50%.

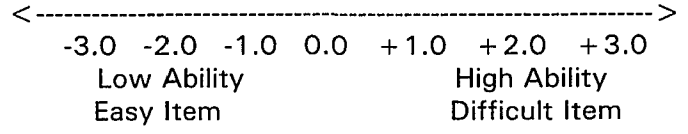
The two-parameter logistic (2-PL) model (Birnbaum, 1968) can be viewed as an extension of the 1-PL model, where items not only differ in difficulty but also in discrimination. The discrimination parameter, a_i, is proportional to the slope of $P(\Theta)$ at the point $\Theta = b_i$. Furthermore, as in the 1-PL model, guessing does not occur in the 2-PL model. The equation for the 2-PL model is given as follows:

$$P_i = P_i(\Theta) = e^{Dai(\Theta-bi)}/(1+e^{Dai(\Theta-bi)})$$

The three-parameter logistic model (3-PL) adds a third parameter, c_i, the pseudo-guessing parameter. The 3-PL model incorporates this parameter to account for the misfit of items at the low end of the ability continuum due to low ability examinees giving a correct response. The 3-PL incorporates the pseudo-guessing parameter as follows:

$$P_i = P_i(\Theta) = c_i + (1-c_i)\{e^{Dai(\Theta-bi)}/[1+e^{Dai(\Theta-bi)}]\}.$$

In all three models, the difficulty of an item is defined on the same scale as ability, and in practice, the range of item difficulty tends to be from -2.0 to +2.0, assuming that ability ranges from -3.0 to +3.0 (Hambleton & Cook, 1977). A value of -3.0 on the scale represents low ability or an easy item, while a value of to +3.0 represents high ability or a difficult item. Examinees who perform well will generally have scores at the higher end of the continuum, while those who perform less well will have estimated abilities at the lower end of the continuum.



Regardless of the model which is utilized, IRT allows one to predict the performance of an examinee, to specify a relationship between the unobservable trait and examinee item performance, and to arrive at an estimate of proficiency for the ability being measured (Hambleton & Swaminathan, 1985).

B. Estimation of Ability in IRT

While the development of IRT helped overcome some of the unresolved issues of classical test theory, its main advantage could be noted in the ability to deal with items one at a time (Wainer, 1990). Items can be arranged from least to most difficult based on an underlying trait. The situation suggests that all examinees need not be presented the same items, but only those which will assist in placing an individual along an ability continuum. Thus the full power of computerized adaptive testing is made available through the variable-branching strategies suggested by IRT (Weiss, 1985). Most current work with CAT is thus based on IRT methodologies. In fact, CAT may be viewed as a practical application of IRT.

Bayesian as well as maximum likelihood procedures may be utilized in obtaining proficiency or ability estimates based on CAT (Hambleton & Swaminathan, 1985). The former is especially useful when prior information regarding the distribution of abilities of a group is available.

Owen's procedure (1975) involves the individually tailored sequential design of a test by appropriate choice among available items and estimation of ability (Θ) via a Bayesian-motivated approximation. At each step of the ability estimation (m), a normal prior distribution is assumed with parameters μ, σ^2. Here, m represents the number of

items already presented, and thus an item to be selected at step $m+1$ is chosen as to maximize information about the newly estimated examinee ability. Owen (1975) demonstrated that the test scores achieved in such an approach is an approximation of the examinee's latent ability, and as the number of items administered increases, this estimator of ability approaches the examinee's true ability.

Alternatively, a maximum likelihood estimation (ML) procedure may be utilized to obtain ability estimated in a CAT situation (Hambleton & Swaminathan, 1985; Thissen, 1982). A maximum likelihood estimate of a person's ability can be obtained from an arbitrary set of items for which continuous response functions with respect to a common dimension can be specified (Birnbaum, 1968). In such a procedure, the likelihood function $L(u'|\Theta)$ given below may be viewed as criterion function, and the value of Θ that maximizes the function can be taken as an estimator of examinee ability.

$$L(u1,u2,u3...un\,|\Theta) = \prod P^{ui}\, Q^{1-ui}$$

In a sense, the ML estimator of Θ may be viewed as the value of the examinee's ability that generates the greatest probability of an examinee's response pattern (Hambleton & Swaminathan, 1985). While a graphical procedure may provide this estimator, it is more common to maximize the natural logarithm of the likelihood function. The maximum of the loglikelihood function $\ln L(u|\Theta)$ is attained when Θ satisfies the equation

$$\frac{d}{d\Theta} \ln L(u|\Theta) = 0.$$

This equation, also known as the likelihood equation, can be solved iteratively through the Newton-Raphson procedure (Hambleton & Swaminathan, 1985), where the mth

approximation of Θ given a response pattern is obtained using the first and second derivatives of the loglikelihood function as follows:

$$\Theta_m = \Theta_0 - \left\{ \frac{d}{d\Theta} \ln L(u|\Theta) \quad / \frac{d^2}{d\Theta^2} \ln L(u|\Theta) \right\}.$$

This process continues until convergence takes place, often taken at the point where the difference between successive approximations of Θ is less than or equal to 0.001.

ML estimation, however, fails in estimating ability when a string of all correct or all incorrect responses are given in a CAT. In such situations, ability estimates will approach positive or negative infinity, respectively (Hambleton & Swaminathan, 1985). A Bayesian approach can be implemented, or alternatively a set of "false" items may be used in this initial stages of the calibration process to mimic the processes of a Bayesian approach in a ML estimation procedure (Halkitis, 1993).

In certain cases, ML procedures have been shown to be preferable to the Bayesian approach (Weiss & McBride, 1984). ML estimation of ability provides less biased estimates when no differential prior information of ability is available or when prior information that is available might be inaccurate.

Recently, EAP (estimated a posteriori) estimation has been specified as a means of ascertaining ability estimates where efficiency of computation is at a premium (Bock & Aitkin, 1981; Bock & Mislevy, 1982). The EAP estimate requires significantly fewer operations than either Owen's Bayes modal (MAP) or maximum likelihood estimators and has demonstrated near equivalence in measures of precision. In the EAP procedure, posterior standard deviations indicate the precision of the ability estimates. The estimation of ability is based on a numerical evaluation of the mean and variance of the posterior distribution, and calculations are conducted noniteratively, simply

summing the log likelihoods as items are administered given a predetermined set of probabilities of correct responses at a fixed number of points. Standard errors of estimate are calculated in ML approaches. In the equation that follows, the estimation of ability is ascertained from the summation of likelihood of a correct response (L_j), the point of estimation, and the predetermined weight associated with that point of estimation, from point 1 to point q.

$$\Theta j = \sum_{1}^{q} X_k L_j(X_k) * W(X_k) / [\sum L_j(X_k) * W(X_k)]$$

The standard error of an ability estimate is based on the notion of information, a quantity inversely proportional to the squared length of the confidence interval around an estimate of an examinee's ability (Birnbaum, 1968). The standard error of ability estimate is equal to the multiplicative inverse of the square root of this information and is statistically a function of an examinee's latent ability (Lord, 1984). The information function is given by the second derivative of the loglikelihood function.

$$I(\Theta) = \frac{d^2}{d\Theta^2} [\ln L(u|\Theta)]$$

When information is high, confidence bands about the estimate of ability tend to be narrow; when the information is low, then confidence bands about the ability estimate are a large. Because maximum likelihood estimators are asymptotically normally distributed, the maximum likelihood estimator of Θ is asymptotically normal with mean Θ and variance $[I(\Theta)]^{-1}$, where $I(\Theta)$ is the information function. The reciprocal of the information function is thus the variance of the maximum likelihood estimator Θ. Thus standard errors of ability estimates are given by

$$(\Theta)]^{-2} = \frac{d^2}{d\Theta^2} [\ln L(u|\Theta)].$$

The standard error of estimate has been suggested to be a major index of dependability when considering a computerized adaptive test (Green et al, 1984; Samejima, 1977). While indices such as reliability have traditionally been considered in assessing the dependability of examinations, such indices have less applicability in CAT situations that utilize IRT methodologies. Since the standard error of estimate is a construct defined separate from any group of subjects being tested, unlike the concept of reliability, and can be considered solely as a property of the test or testing procedure then it should be considered as fundamental to the issue of dependability and accuracy in test theory using a CAT approach. Consequently, it is suggested that testing consumers attend to standard error of measurement rather than reliability in evaluating test effectiveness (Green et al, 1984).

C. The Nature of Adaptive Testing

The psychometric framework of IRT and procedures for ability and precision estimation provide the mathematical underpinning of CAT. Yet CAT is more than simply an application of IRT; it is instead a marriage of IRT with the theoretical orientation of adaptive testing that is manifested through the interactive administration by microcomputer. CAT is a system for administering tests in which a computer terminal is utilized to administer items and in which adaptive (tailored) sequencing of items, rather than lock-step sequencing of conventional tests, is used (McBride & Sympson, 1982). The idea of sequentially tailoring test difficulty to the examinee's ability is at the heart of this testing. The impetus for adaptive or tailored testing stems from the notion that an examinee is measured most efficiently when test items are

neither too easy nor too difficult for him or her (Lord, 1980).

In a traditional or conventional paper-and-pencil examination, the test constructor is limited in the number of items that can be administered and is further hampered if the purpose for which the test is designed is to measure a wide variety of trait levels (Weiss, 1985). Given these limitations, most conventional tests are either peaked or rectangular in their measurement precision.

In a peaked conventional test, the items which are chosen are highly discriminating but possess difficulties centered about a narrow range. The result of such a test construction is an examination that has maximum information about examinees who fall in that narrow ability range, but provides much less information or precision for examinees further away from the peak. In effect, an information curve that peaks and is high at a narrow range of ability.

In a rectangular conventional test, items are chosen so that the difficulty spans equally across all levels. Equidiscriminating items having a wide range of difficulties are thus incorporated. The problem encountered here is that examinees will tend to encounter items that are either too difficult or too easy for them. While a rectangular conventional examination will provide relatively equal levels in precision of measurement at all ability levels, the overall magnitude of the precision in measurement will be relatively low (Weiss, 1985). In effect, a relatively flat but low information curve is achieved.

The ideal situation is one in which a flat yet high information curve is achieved (McBride & Weiss, 1984). The only way to achieve this end is to administer to each examinee a different subset of items that will provide the most information about his or her latent ability. The result is a series of peaked distributions for the examinees

which translates to a high and flat information curve for the entire examinee pool. The item pool provides for a family of multiplexed tests from which to choose (Kingsbury & Houser, 1993).

In an adaptive test, this situation is realized by selecting items from a pool so that each examinee encounters an examination consisting of items peaked at his or her ability level. As a result, an adaptive test provides equal measurement precision for all examinees along an ability continuum. An adaptive test selects items during the course of testing in such a way as to maximize the information for each examinee (Weiss & McBride, 1984). This is accomplished by selecting items sequentially on the basis of an examinee's performance (McBride, 1979). By selecting items in this manner, the information function for each examinee taking the test is maximized, and as a result, a smaller standard error of estimate of ability or greater precision in estimation of ability may be ascertained.

D. The Implementation of CAT

Like any adaptive or tailored test, a computerized adaptive test is a methodological framework that can be implemented to assess an examinee's ability in a given area. Almost every application of CAT in the last fifteen years has benefited from and depended on IRT (Kingsbury & Houser, 1993).

The CAT procedure usually consists of the following components: (1) an IRT model (2) an item pool (3) an entry level (4) an item selection procedure (5) an estimation or scoring method and (6) a termination or set of termination criteria (Weiss & Kingsbury, 1984). If a test which is computerized adaptive, items are selected for

administration based on the responses of items previously administered (Stocking, 1987). Different sets of items are administered to individuals depending on the individuals' level of the trait being measured (Weiss, 1985). The computerized aspect of the CAT stems from the fact that items are selected and administered by a computer; the adaptive characteristic stems form the fact that items are selected to provide maximal information about the examinee.

In such examinations, assuming that we have no information about the examinee being measured, an item of average difficulty is administered first. An initial Bayesian estimate of ability may be provided by awarding each examinee one success and one failure on two dummy items of mean item difficulty (Halkitis, 1993). If the examinee answers correctly, then a more difficult item is presented; if the answer is incorrect, then an easier item is presented (Lord, 1980). After each response, a revised estimate of the examinee's ability is ascertained (Urry, 1977). Lord (1970) used the Birnbaum model to select items in an adaptive process. In his approach, the selection of the "next" item was based on the estimate of the examinee's ability as well as parameters of items from a precalibrated pool.

A CAT test continues in this fashion, making greater jumps in ability estimates at first but then stabilizing as more items are presented, and is terminated when one of a set of criteria have been satisfied. Such stopping rules include a preselected number of items being administered, a predetermined amount of time elapsing, or a given level of precision being attained (Thissen & Mislevy, 1990; Urry, 1977; Weiss, 1985).

Both Owens' Bayesian approach (1975) and maximum information approach (1982) may be utilized in item selection (Kingsbury & Zara, 1989). Owen's (1975)

Bayesian adaptive testing strategy estimates Θ after each item response, and then makes use of the "unused" item which is most informative at the given ability level (Weiss & McBride, 1984). The procedure involves the selection of items that minimizes the variance of the posterior distribution of examinee ability. As the number of items administered increases, the posterior distribution becomes more concentrated and reflects the increased accuracy with which the examinee's ability is estimated. The effectiveness of this approach has been demonstrated in simulated testing situations (Urry, 1971) and using the actual responses of live examinees to 598 mathematics items drawn from four conventional precollege tests taken at full length by examinees (Jensema, 1972). Yet as was previously noted, the success of this approach depends, in part, in the appropriateness of the prior distribution of abilities.

The maximum information approach (1982) based on maximum likelihood estimators involves the election of items that provide maximum information, and which in turn most greatly reduce the standard error of estimate (Hambleton, Swaminathan, & Rogers, 1991). This thinking is in line with Wainer's (1989) suggestion that the function of an adaptive test is to mimic automatically what a wise examiner would do in the testing process. Thus, when an examiner presents an item that is too difficult, the next item would be easier because we would learn less about an individual's ability from items that are too difficult or too easy than from items that match the examinee's proficiency. This in sharp contrast with conventional fixed item tests where all individuals are administered the same item set.

In the maximum information approach, the selection of items for a CAT is based on the item information index, a statistic which describes how precisely an item measures at various points along the continuum (Weiss, 1985). The item information

curve associated with each item has its peak at the difficulty of the item, and the height of its peak is related to the item discrimination. Information functions for items based on specific algorithms relevant to each of the three logistic models depend on the item response function and the conditional variance at each ability level. These functions are generally bell-shaped; maximum information for the 1-PL and 2-PL models are obtained at b_i, the difficulty of the item (Hambleton & Swaminathan, 1985). For the 3-PL model, maximum information is given at the ability level Θ given by

$$b_i + 1/Da_i \ln[1/2 + 1/2*(1 + 8c_i)^{.5}],$$

where a_i, b_i, c_i represent the parameters difficulty, discrimination, and pseudo-guessing.

In its most basic sense, CAT requires a computer to present each item, score the response, and then select the next item for administration to the examinee (Green et al, 1984). The procedures for estimating ability and item selection as noted above have provided the workings of several models testing the effectiveness of the computerized adaptive approach. While early attempts at adaptive testing (Angoff and Huddleston 1958; Cleary, Linn, & Rock, 1968; Lord, 1970; Wright & Douglas, 1975), were based on modified versions of printed examinations or theoretically driven by classical test theory, recent efforts have been based on IRT-based procedures for ability estimation and item selection.

In the end, CAT has been suggested to demonstrate several advantages of conventional testing procedures (Weiss, 1985). And while the main research with CAT has focused on ability and not achievement testing, efforts to document the effectiveness of the approach have suggested, for the most part, that CAT has yielded measurements of comparable or superior quality to those of conventional tests while making use of relatively fewer items administered to each individual. These

improvements are in terms of greater precision of measurement for all or most trait levels (Weiss, 1985).

Yet the main ingredient for a good CAT is a large, well-distributed item pool with well estimated item parameters, for a domain with one dominant dimension (Green et al, 1984; Wainer, 1993). Urry (1977) refers to them simply as excellent item banks; Hambleton, Swaminathan, and Rogers (1991) claim that any item bank for a CAT would consist of hundreds, and possibly thousands of test items. The following section considers the issue of precision of measurement in light of the item pools that were used to generate the findings of the previous research.

E. CAT and Precision in Measurement

Along with improvements in efficiency (Stocking, 1987; Urry, 1977; Wainer, 1989; Ward, 1985; Weiss, 1985; Weiss & Kingsbury, 1984), a CAT is suggested to provide improved measurement characteristics (Wainer, 1989; Ward, 1985; Weiss, 1985). CAT can reduce the length of many tests by as much as 50% without a loss of measurement precision (McBride & Sympson, 1982). These improvements have not only been demonstrated in models using the dichotomous item repose model, but also on examinations using the graded response model (Dodd, Koch, & De Ayala, 1989), as well as on tests determining mastery classification (Lewis & Sheehan, 1990; Sheehan & Lewis, 1992; Weiss & Kingsbury, 1984). It has been argued that it is the efficiency of CAT that has attracted the interest of psychometricians (McBride & Martin, 1983). Yet the issue of precision is one that has also been noted.

These improvements in measurement are manifested in greater precision of

ability estimation for all or most trait levels, which translate into higher levels of reliability and potentially higher levels of validity (Weiss, 1985). In addition, examinations presented in a CAT format have been shown to be comparable in the traits that they measure as their conventional paper-and-pencil counterparts (Henly et al., 1989; Cudeck, 1985; Moreno et al., 1984). The findings of these studies suggest that some degree of structural equivalence can be expected when conventional measures of differential abilities are presented in a carefully developed adaptive mode (Henly et al., 1989).

The issue of precision has been examined extensively over the last decade using both simulation and live-testing situations to generate data. While certain common characteristics appear throughout much of the research, different types of item pools, stopping criteria, and estimation procedures have yielded results that when taken together might provide further insight into the issue of precision of measurement in CAT.

Using data on the Armed Services Vocational Aptitude Battery (ASVAB), Divgi (1989) utilized the indices of multiple correlation and communality to suggest that on all but two subtests, the CAT version of the ASVAB possessed a higher reliability than the paper-and-pencil versions of the subtests. Data for the analysis were ascertained from a validity study (Vicino & Hardwicke, 1984) of the CAT version of the ASVAB in which recruits took a CAT version of the examination plus three to five subtests of the paper-and pencil-version. CAT scores were obtained using Owen's (1975) procedure and then equated to the paper-and-pencil metric. Using a series of covariates the multiple correlation of each version with the covariates were computed; the ratio of CAT and paper-and-pencil were determined by regressing both scores on the available

covariates, indicating that on the subtest of General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, and Electronics Information, The CAT version of the ASVAB possessed a higher reliability. Yet these results have not always been corroborated (Bryson, 1971).

Urry (1977) argues that holding test length and all else constant, a good adaptive test is superior to a conventional test provided that highly discriminating items are available. In fact, it was reported that for a sample of 57 Civil Service job applicants, an adaptive verbal test of the Civil Service Commission's Professional and Administrative Career Examination (PACE) achieved an 80% reduction in the test length required to attain any of several prespecified levels of reliability (Urry, 1977).

In many cases Monte Carlo computer simulation studies have been utilized to examine the issue of precision. In such a procedure, the item response model is used to administer hypothetical tests to hypothetical examinees. Since the examinees (simulees) are generated by the computer, they may possess any desired distribution of trait level and can respond in prespecified ways to items which are presented (Urry, 1970; Weiss, 1985).

The findings of such simulation studies (Maurelli & Weiss, 1981; McBride, 1977; Weiss & McBride, 1984), have noted that adaptive tests measure with not only greater efficiency but also with a greater precision as compared to conventional fixed-length examinations.

McBride (1977) reports on a series of studies to examine the feasibility of CAT using a Bayesian adaptive ability testing strategy. In one of the studies, the effects of guessing and item pool characteristics were considered in light of the trait estimates achieved using Owen's Bayesian sequential testing strategy (Owen ,1975). Two ideal

item pools were simulated. Both pools were assumed to have a distribution that ranged from -2.4 to +2.4 logits, with pseudo guessing parameters of all items set at 0.8. The two pools varied in discriminating power; the first contained items with a discrimination constant equal to 0.9, the second with a constant equal to 1.6. Bayesian estimation procedures were employed and examinee performance were generated via a Monte Carlo simulations, incorporating a stopping rule based on a standard error of estimate of 0.25 and a maximum test length of 30 items. Initial examinee ability was set at $\Theta = 0.0$, and the prior distribution was assumed to be normal (0,1). Despite the ideal nature of the item pool, the precision level was never reached for the pool with lower discriminating items. For the second pool, test length ranged from 12 to 30 items and correlated 0.85 with the ability estimator. Thus while use of the second pool corroborates the notion of increased precision within a more efficient approach, the level of precision ascertained in such an approach is dependent on ability level, suggesting a potential problem in precision achieved if a group is of higher ability than originally assumed. Furthermore, the generalizability of the results is limited to item pools with a composition as described above.

Followup investigations (McBride, 1977) suggested that tendency to overestimate high abilities and underestimate low abilities is characteristic of a CAT approach regardless any of three item pools that are utilized and suggests that the issue of loss of accuracy or precision at the extremes of the ability continuum undermines one major advantage of CAT over conventional tests, the proposed superiority of measurement accuracy at those extremes.

This issue of precision has been corroborated (Stocking, 1987). Using a precalibrated (3-PL) 120 item pool, the performance of 20 examinees ranging in ability

was simulated. A minimum of ten items and a maximum of 40 items was administered to each of 200 examinees who were simulated at nine true score levels. Testing also was terminated when a prespecified standard error of estimate was achieved. In a second condition, a twenty item fixed length adaptive test was simulated. While no clear conclusions were evidenced, it was suggested that the administration of more items generally improves estimation of ability. In addition, it appears that simulees at the extreme of the ability continuum are either overestimated or underestimated in their proficiency estimate.

It may be suggested that the overestimation at the upper ability and underestimation at the lower ability is in fact an artifice induced by the presentation of the final item in the test administration. Perhaps, it is highly probable that the high ability examinee will answer the final item correctly, thus inflating the ability estimate; the low ability individual has a high probability of answering the last item incorrectly, thus leading to an even lower ability estimate. Removal of the final item from the ability estimation process could lead to the amelioration of this problem.

However, in direct comparisons of CAT with tailored peaked conventional tests, the advantages of CAT were noted (Stocking, 1987). Using 120 items from an arithmetic placement test calibrated on the 3-PL model, Stocking simulated the ability of 200 examinees at each of seven true score levels taking a 20 item adaptive test. Results of this testing (RMSE) were compared to the results (SEM) of an actual twenty-item testing. In addition, these results were examined in light of six twenty-item tailored conventional examinations that peaked at a different level along the distribution. In all cases, in terms of precision the adaptive test outperformed the actual test, and in all but three of the comparisons with the tailored conventional test

the adaptive test achieved greater precision.

Using an ideal item pool consisting of an infinite number of items based on the 3-PL model, McBride & Weiss (1984) examined the performance of examinees (simulees) and precision of measurement associated with performance under different conditions for eight data sets. Results suggested that despite the function of adaptive testing to provide equally precise measures along the ability continuum, the realistic conditions of a constant prior Θ estimate (i.e. all ability estimation begins with the assumption that all examinees have the same ability) does not achieve this desired goal when using a Bayesian adaptive testing strategy. It is suggested that a variable prior might lead to equiprecision in measurement along the entire ability continuum. Thus Owen's (1975) adaptive testing strategy yield unbiased estimates of equal precision throughout the ability range only under the completely unrealistic condition of an accurate prior Θ estimate.

Equal levels of precision have also been shown to be obtained readily using a CAT approach (Urry, 1971). Using two ideal item banks consisting of equidiscriminating ($a = 1.6$) items, termination based on precision was obtained with an average of approximately 11 items. The first bank consisted of 20 items with difficulties at each of 5 levels of the Θ continuum and the second consisted of 5 items with difficulties at each of 20 levels along the Θ continuum. In addition, ability estimates were found to correlate highly ($r = .926$, $r = .919$ respectively) with abilities drawn randomly from a population assumed to be normal (0,1).

In direct comparisons with conventional tests, CAT procedures have been shown to attain the same level of precision using half the number of items (Moreno et al, 1984). The relationship between CAT and paper-and-pencil versions of the ASVAB

were investigated using Marine recruits as subjects. Subtests were administered the initial ASVAB before enlisting in the armed services, and were retested using an alternate form ASVAB approximately two weeks after active duty was initiated. CAT versions of the ASVAB were administered to available recruits approximately 24 hours after arrival at the recruit depot. Item pools for the CAT consisted of 225 items in Arithmetic Reasoning, 39 items in Word Knowledge, and 25 items in Paragraph Comprehension. Because the Arithmetic pool was deficient in easy items, 77 additional items were added to the original pool of 148. Items were selected and scored based on Owen's (1975) adaptive testing strategy. CAT score were found to correlate as highly with original ASVAB scores as did the retest even though the CAT subtests consisted of approximately half the number of items. In addition, it was concluded that CAT versions of the ASVAB can achieve the same level of precision as conventional tests while using half the number of items. It has been suggested that paper-and-pencil ASVAB scores are only precise at mid-range ability and less precise at the extremes, but CAT scores will be more precise at the extremes and just as precise at middle range abilities (McBride & Sympson, 1982). These findings regarding the ASVAB have been replicated (Kiley, Zara, & Weiss, 1983; McBride & Martin, 1983) and at least support the contention that the efficiency and accuracy of CAT are applicable in live-testing situations.

The advantages of CAT versions of the ASVAB were also noted by McBride (1980) who compared the reliabilities of fixed length adaptive tests, variable length adaptive tests, and conventional tests consisting of different lengths. Reliability, which was defined as the correlation between scores on alternate forms of a given test length, was found to be substantially higher for either version of the adaptive test than

for the conventional test. More specifically, the reliability of a 5-item adaptive test was found to be equivalent to that of a 15-item conventional test.

In another comparison of conventional and computerized adaptive testing strategies, Johnson and Weiss (1980) examined measurement precision under live testing situations using college students. An item pool of 256 vocabulary items were parameterized using the 3-PL model and assumed to have a guessing parameter equal to .20. The 60 most informative items at $\Theta = 0.0$ were used to compose the alternate forms of the conventional test. Both Bayesian and maximum likelihood methods were used to achieve ability estimates and errors of measurement, posterior variance and standard error of estimate respectively. Along the entire ability continuum, both adaptive testing strategies yielded smaller standard errors of measurement than the conventional test. However, as was the case with the conventional test, the adaptive tests measured less precisely at the extreme ability levels that they did about the mean.

Thissen (1990) examined the precision of a hypothetical computerized adaptive test, the GCAT, by comparing it to the paper-and-pencil version. In assessing the precision of the GCAT, information curves were ascertained for 2000 simulees on each of two forms of the GCAT as well as for the paper-and-pencil version on each of the examination subtests. In all cases, the information curve of the CAT was found to exceed that obtained from the paper-and-pencil administration, even though the conventional examinations consisted of more items.

Early work using a fixed branching approach to item section has also demonstrated the superiority of measurement precision for a CAT (Lord, 1980; Lord, 1977). Approximately 1000 examinees were simulated taking a broad-ranged test of

verbal ability in which testing stopped after 25 items had been administered, corresponding to a type of blueprint of item difficulties. Because the test was designed to measure verbal ability throughout the various school grade levels, the items were drawn from a variety of sources including *The Cooperative School and College Ability Test, Cooperative Sequential Tests of Educational Progress, The PSAT, SAT,* and *GRE.* An initial pool of 900 items was narrowed to 363, retaining all items at the extremes due to the scarcity of these types, and then randomly selecting items from the intermediate levels. In the end, 25 items were available at each of the ten difficulty levels, excluding the extremes.

Item difficulties were arranged in rows and entry level was determined by the grade level or "some other rough estimate" of examinee ability. If the examinee answered correctly the first item, then he or she would be presented a more difficult item from the second row; if the first answer was incorrect, then an easier item from the second row was selected. ML estimates were ascertained after each response.

In direct comparisons to three forms of the conventionally administered *Preliminary Scholastic Aptitude Test*, it was found that the adaptive test was at least twice as good in the information obtained for examinees. To allow for direct comparison with the 25 item CAT, PSAT Verbal scores were adjusted to 25 items. In comparisons of the information function, the CAT provided significantly more information along the ability continuum than any of three paper-and-pencil examinations. While both types of examination provided maximal information at an ability level of approximately 0.75, the CAT provided an information value of approximately 60 (standard error of estimate = .13), while the paper-and-pencil version provided an information value of approximately 20 (standard error of estimate = .22).

Although the CAT demonstrated superiority in precision of measurement to the conventional test, a CAT consisting of 363 items also was found to be at least twice as good as a 182-item CAT, suggesting the need for large item pools that span ability. Lord (1980; 1977) suggests that doubling the size of the item pool of the CAT will give a much better test because selecting the best items from a 363-item pool gives a better set of 25 items than selecting from a 183-item pool.

Chapter III

METHOD

A. Rationale for Research

While there has been a substantial amount of research in simulated situations and some live-testing situations, issues regarding CAT require further elucidation. Perhaps one overriding question concerns the impact of the item pool that is utilized on the effectiveness of a computerized adaptive test. How are indices of accuracy a function of the items that constitute a CAT? In fact, findings in some investigations (McBride, 1977) have cautioned about the generality of results regarding CAT accuracy, noting that indices of accuracy may be limited to "ideal" item pools with rectangular distributions.

The application of the more efficient, and often shorter length, CAT is not without its own problems. Since tests presented in a CAT format are shorter, estimates are more vulnerable to the idiosyncracies of item performance and thus item pools of substantially greater sizes might be required (Johnson & Weiss, 1980; Wainer & Kiley, 1987). Standard error of estimate curves have been shown to be substantially better when test length is increased (Hambleton & Cook, 1983). Because the score information curve (information function) of a CAT based on IRT methodologies is dependent on the parameter estimates of the item pool and this information function is directly linked to the standard errors of estimate, practical concerns arise about the precision of ability estimates generated in this manner (i.e. parameter estimates should be accurate), suggesting perhaps that a large item pool is necessitated.

While no specific guidelines exist for the appropriate size and characteristics of item pools, it has been suggested that 100 items might provide satisfactory results for a CAT so long as the item difficulties span the full range of trait levels in the population and items possess high discrimination (Weiss & Kingsbury, 1984), but 150-200 items would be better (Weiss, 1985).

Others (De Ayala, Dodd & Koch, 1990) have noted that the item pool of a CAT requires that the examiner have access to a pool of 200 to 500 items and a database of responses to the items by examinees ranging in number from 300 to 1000; yet without items that span the entire difficulty continuum, increases in measurement quality beyond that of conventional test cannot be assured.

The characteristics of the item pool that is utilized in a CAT will ultimately affect the precision of the estimates of ability that are ascertained. This issue has been noted in some of the studies cited above, yet a direct comparison of precision between conventional tests and CAT, limited by the realities of an item pool, might help to clarify the issue and shed light on the precision of CAT in most achievement testing environments where ideal situations are not possible.

B. Overview of Procedure

This comparative and evaluative investigation aimed to examine the precision of measurement of a computerized adaptive examination using an item pool limited in its composition. The investigation sought to determine the extent to which the precision of measurement of examinee ability was enhanced when items, constituting a limited item pool, were presented in a CAT format as compared to a paper-and-pencil

administration in which groups of examinees were presented the same set of items. What is the effect on the precision of ability measurement when an examination with a limited item pool is administered in a CAT format as compared to a paper-and-pencil, non-adaptive framework? Because the ability of CAT to enhance accuracy rests on maximizing the information about the entire ability continuum by presenting examinees maximally informative items, this advantage was tested in a situation where maximally informative items were not available throughout the testing process given the limited nature of the item pool. The hypothesis was tested using an item pool considered to be typical of many commercial achievement testing situations.

To assess the advantage in precision of CAT administration in a situation when an item pool is limited in nature, measures of precision were ascertained for 4494 examinees on one of four paper-and-pencil versions of an examination where items were randomly assigned to each of the four versions. Indices of precision were compared to those achieved when a tailored set of items were administered to a simulated examinee pool. Standard errors of measurement ascertained in the paper-and-pencil administrations were compared to those ascertained in a simulated CAT situation where a fixed number of items were administered. Further, the number of items required for the CAT with this limited item pool to achieve the same level of precision of the paper-and-pencil administrations was determined. The goal of these analyses was to explore the differences in precision achieved via these two testing frameworks given the limited nature of the item pool that was utilized.

Ability estimates and measures of precision were ascertained for a randomly selected sample of 4494 examinees who were administered one of four versions of *Calculations of Drug Administration*, a pilot achievement examination designed for

registered nursing students who had completed the appropriate course preparation in pharmacology principles at accredited schools at the time of test administration in February, 1992. Examinees were students in randomly selected schools of nursing in the United States. Four versions of the examination each consisting of 30 unique items and linked by a set of fifteen anchor items were utilized to allow for the experimentation of a larger set of items.

Examinee responses were calibrated for the 1-PL model using *BILOG 3* (Mislevy & Bock,1990). Initial calibrations of item difficulty and fit statistics were examined to assess the overall fit of the items to the model. Items judged as misfitting were then eliminated from the item pool (Halkitis, 1992; Hambleton, Swaminathan & Rogers, 1991).

C. Paper-and-Pencil Administrations

The final pool of 101 items were recalibrated using *BILOG 3*. Indices of fit of items were assessed to assure that the items fit the Rasch model. Mean-square statistics were computed for the items; these indices of fit in conjunction with the power of the hypothesis test were used to judge items as either fitting or misfitting (Halkitis, 1992). In addition, estimates of ability and standard errors of measurement were calibrated for each examinee pool. The final item pool utilized in these calibrations consisted of twelve anchor items and 89 unique items. The four versions of the examination consisted of 33, 33, 35, and 36 items respectively, linked by a set of twelve common items. The ability estimates, standard errors of measurement, test, and test length provided the basis of comparison with the CAT simulated data.

D. Simulated CAT Administration

The 101 calibrated items served as the CAT item pool which was utilized to simulate examine abilities and standard errors. Simulees were drawn from a hypothetical uniform distribution (-3.00, +3.00) at twelve critical points (-1.00 to +1.75 inclusive, at intervals of 0.25 logits along the continuum). The critical range was chosen as it represented the ability continuum ascertained in the paper-and-pencil administrations. Fifty simulees were generated at each of the critical points along the continuum to simulate an examinee pool of 600 students. For each, a maximum of 36 responses were generated so to adhere to the maximum test length of the lengthiest paper-and-pencil administration. Ability estimates and standard errors of measurement were noted after the adaptive administration of each item. This data provided the basis for comparisons of precision at the critical points in each of the four traditional test administrations.

E. The Generation of Simulee Abilities

Responses were simulated using a random number generator written in the programming language *C++* (Borland, 1993). After the determination of the next item to be administered based on maximizing information (i.e. the item with difficulty that most closely matched ability estimate), $P(x = 1 \mid \Theta)$ was calculated; Θ was taken as the latent ability the simulee was assigned. To determine if the response was correct or incorrect, a random number was generated; if this number fell within the probability of $P(x = 1/\Theta)$, then the response was determined to be correct; else, it was incorrect (see

Appendix A). This response was entered into a *Paradox v. 4.0* (Borland, 1990) database where the new estimated ability and standard error were calculated using a ML approach (see Appendix B). The same procedure was undertaken for the administration of each of the 36 items where the presumed latent ability of the examinee in conjunction with the item difficulty was used to determine the probability of correct response, and ML procedures were used to estimate ability of the simulees.

To initiate the estimation of ability and standard error and to allow for the convergence of the ML estimates, two responses (one correct, one incorrect for two items) were presupposed for each simulee. The items were each assumed to have a difficulty parameter of 0.00 resulting in an initial ability estimate of for each simulee. Given no prior information, 0.0 presented a "best-guess" estimate of ability for the presupposed uniform distribution of -3.00 to +3.00. The first '"true" item presented was of difficulty 0.01 as this difficulty was closest to the presumed average ability of the simulee pool. Three such items were part of the 101 item pool.

The step by step adaptive procedure utilized was as follows:

(1) Set the initial estimate of Θ equal to 0.0 by assuming the presentation of two false items of difficulty 0.0, one marked as correct response, the other as incorrect response. These two false items were dropped from the estimation after the administration of fifteen "true" items.

(2) Present an item from the pool closest to 0.0 logits; in this case an item of difficulty 0.1 logits.

(3) Flag the item as presented, so that it will not be administered again to the same examinee.

(4) Determine the probability of a correct response given examinee estimated ability

and item difficulty.

(5) Randomly generate a number from .01 to 1.00.

(6) If randomly generated number is less than or equal to the probability of correct response, mark the response as correct; else mark the response as incorrect.

(7) Enter response, as either correct or incorrect, into Paradox database and recompute ability estimate.

(8) Select as the next item for presentation the item whose difficulty is the smallest absolute distance from newly estimated ability. Absolute distance is measured as the difference in logit units between the item difficulty and estimated ability.

(9) Repeat steps 3 through 7 until 36 items are presented.

F. Comparing Testing Frameworks

In assessing the advantages of a CAT methodology, the data were examined using two approaches. In the first analysis, the precisions of measurement ascertained in the CAT administration of a fixed number of items were compared to the standard errors of measurement obtained in each of the fixed length paper-and-pencil administration at each of the thirteen ability points along the continuum. Comparisons of the standard errors of measurement at these fixed points provided a method for assessing he precision of the two approaches after a given numbers of items were administered and allowed a judgment to be made regarding the precision that is achieved. In addition, the non-parametric Sign test provided a statistical indication for comparing the accuracy achieved by the two frameworks.

In the second analysis, the number of items required by the CAT procedure to

achieve the level of precision ascertained in the paper-and-pencil administration was computed. Comparisons of the number of items required to achieve this equiprecision provided a basis for assessing the advantages of the CAT administration.

For the purposes of the comparisons, examinees answering all questions correctly or incorrectly on the paper-and-pencil administrations were not utilized as ML estimation would provide less than ideal measures for these examinees. Further, estimations for the CAT simulation incorporated an assumption of at least one incorrect response at mean ability level to initiate calibration. The final comparison group was composed of 1185, 992, 1097, and 1097 (N = 4371) examinees on Forms A to D respectively (See Table 1). As indicated, ability estimates range from a low of -0.98 on Form B to a maximum ability estimate of 1.79 on Form C. For that reason, the critical points of comparison were chosen to range from -1.00 to +1.75 from a uniform distribution of simulees.

Table 1

Summary Statistics for Performance
Four Versions of Paper-and-Pencil Administration
in Calculations of Drug Administration

Form	N	Mean Estimated Θ	S.D Θ	Min Θ	Max Θ	Mean Raw	S.D Raw
A	1185	0.86	0.50	-0.84	1.66	25.67	5.03
B	992	0.87	0.54	-0.98	1.65	26.21	5.08
C	1097	0.87	0.54	-0.88	1.79	26.82	5.49
D	1097	0.87	0.50	-0.63	1.74	27.93	5.25

Because the investigation assumed the prior distribution of examinee ability to be uniform in nature, further exploration considered findings in light of non-uniform distributions. The data generated from the uniformly distributed group provided the basis for approximating estimates when the distributions of simulees were assumed not to be uniform. Five normal distributions were considered to provide a check on the functioning of the CAT with other types of distributions. The following five non-uniform distributions were considered: N(0,1), N(0.5,1), N(-0.5,1), N(1,1), and N(-1,1), where N refers to a normal distribution and the values represent the mean and standard deviation of the distributions, respectively. These analyses were conducted by creating various distributions of the simulated examinees. A check on the mean standard error of estimate after a fixed number of items were determined by taking weighted means of the uniform distribution. Similarly, the mean number of items needed to achieve a given level of precision were determined for the normal distributions by taking weighted means of the number of items needed to a given level of precision in the uniform distribution.

The final item pool of the *Calculations of Drug Administration* achievement examination consisted of 101 items which were calibrated using the Rasch model. The twelve anchor items that were presented in all four versions of the paper-and-pencil administration were used.

Table 2 provides the both Rasch and classical summary item statistics for the item pool which provided the basis of the CAT administration. Mean item difficulty for the 101 item pool was calibrated at -0.143 logits with a standard deviation of 0.566. Maximum and minimum item difficulties were 1.711 and -1.362, respectively, providing a range of 3.073 logits. A majority of the items were centered about 0.0

logits, with a mode of -0.77 and median value of -0.228, constituting a slightly positive skew (skewness = 0.584) distribution (see Figure 1). These results translate to a mean p-value of 0.79 with a standard deviation of 0.14.

Summary item statistics for the four paper-and-pencil examinations are provided in Table 3. Forms A, C, and D (33 35, and 36 items respectively) consisted of a pools with mean difficulty close to 0.1 logits; version D consisted of items with mean approximately -0.2 logits. Of all four versions, the 36-item version C provides the widest range of item difficulties spanning from -1.133 to 1.711 logits; Form A provides the narrowest difficulty band, ranging form -0.787 to 0.919 logits. Modes for forms A through D are -0.787, 0.770, -0.601, and -0.339 respectively, and suggest the ability range at which maximum information is provided. All four distributions depict a positive skew suggesting a lack of difficult items (See Figures 2 through 5).

Figures 6 though 9 provide depictions of the information function for the four paper-and-pencil forms of the examination. Standard errors for the critical points in the distribution are provided in Table 5. Form A, which consists of the narrowest band of item difficulties, also is associated with the largest range of standard errors. Form C, the form that includes the most difficult item also has the smallest standard error at the extreme positive critical point (b = +2.00, s.e. = 0.38), and Form B, which includes the least difficult item, has the smallest standard error at the extreme negative critical point (b = -1.00, s.e. = 0.23).

Precision estimates for the paper-and-pencil administrations are provided in Table 4 for each of the twelve critical ability points after the administration of 33 (Forms A, B) 35 (Form C), and 36 (Form D) items. At each of the ability levels, SEMs for 33 administered items differ no more than 0.3 SEMS from the SEMs achieved on a 36-

item paper-and-pencil examination. Further, SEMs are lower for Forms B and C at each of the ability levels due to the fact that accuracy is enhanced when test length is increased. These SEMs provided the basis of comparison with the CAT simulation.

Table 2

Summary Statistics of Item Difficulty for 101 Items Used in CAT Administration of Calculations of Drug Administration

N	Man Logit	S.D Logit	Min Logit	Max Logit	Percentiles 10	50	90	Mean p-value	S.D p-value
101	-0.143	0.566	-1.362	1.711	-0.79	-0.23	0.68	0.79	0.14

Figure 1
Item Difficulty Distribution of All Calculations Items (N=101)

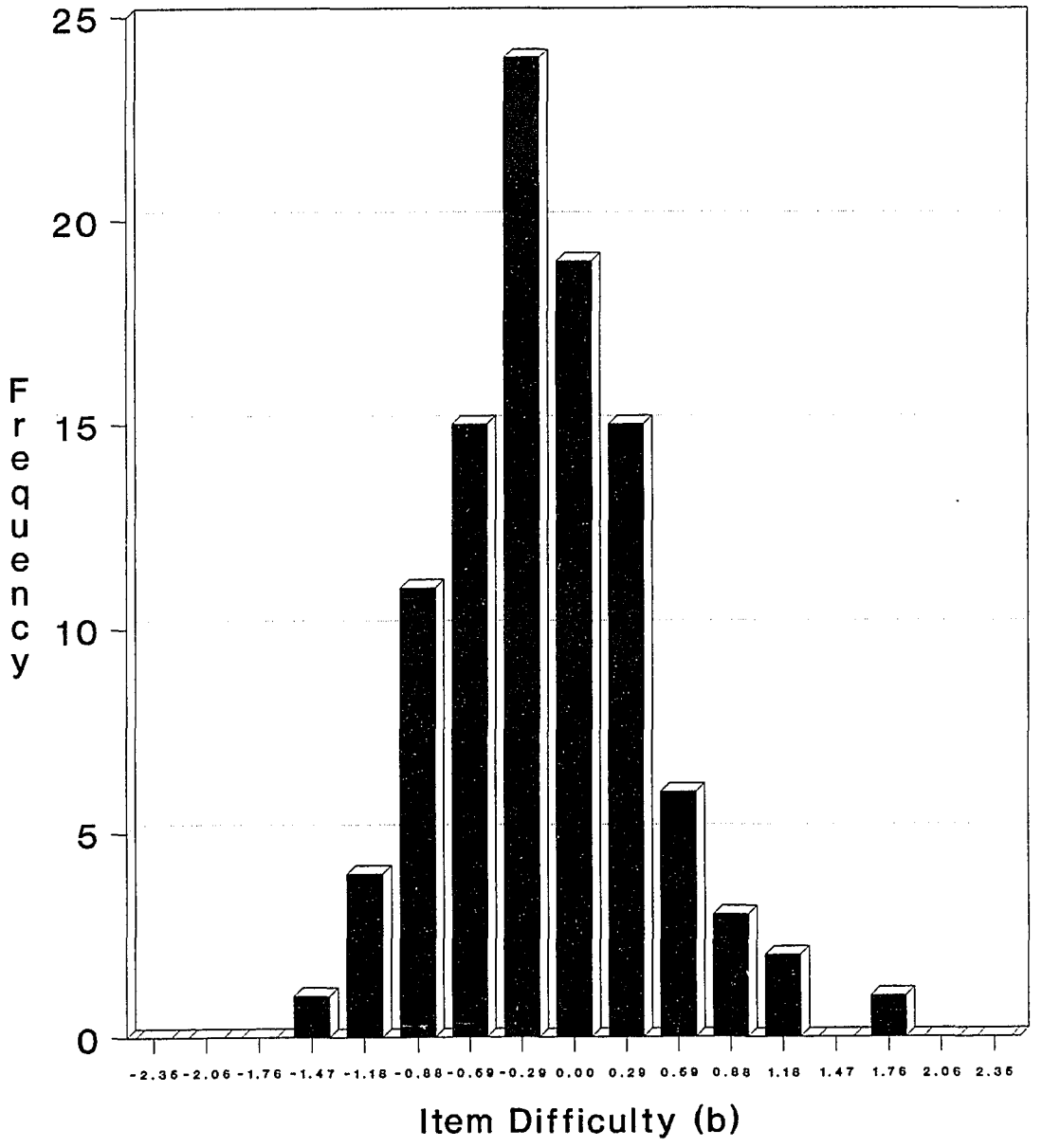


Table 3

Summary Statistics of Item Difficulty for Items Used in Paper-and-Pencil Administration of Calculations of Drug Administration

Form	N	Mean Logit	S.D Logit	Min Logit	Max Logit	Percentiles 10	50	90	Mean p-value	S.D p-value
A	33	-0.063	0.389	-0.787	0.919	-0.62	-0.09	0.40	0.78	0.14
B	33	-0.186	0.551	-1.362	1.284	-0.77	-0.28	0.53	0.80	0.13
C	35	-0.089	0.626	-1.133	1.711	-0.87	-0.13	0.65	0.77	0.16
D	36	-0.086	0.535	-1.206	1.005	-0.78	-0.16	0.70	0.78	0.13

Figure 2
Item Difficulty Distribution
Calculations Examination A (33 Items)

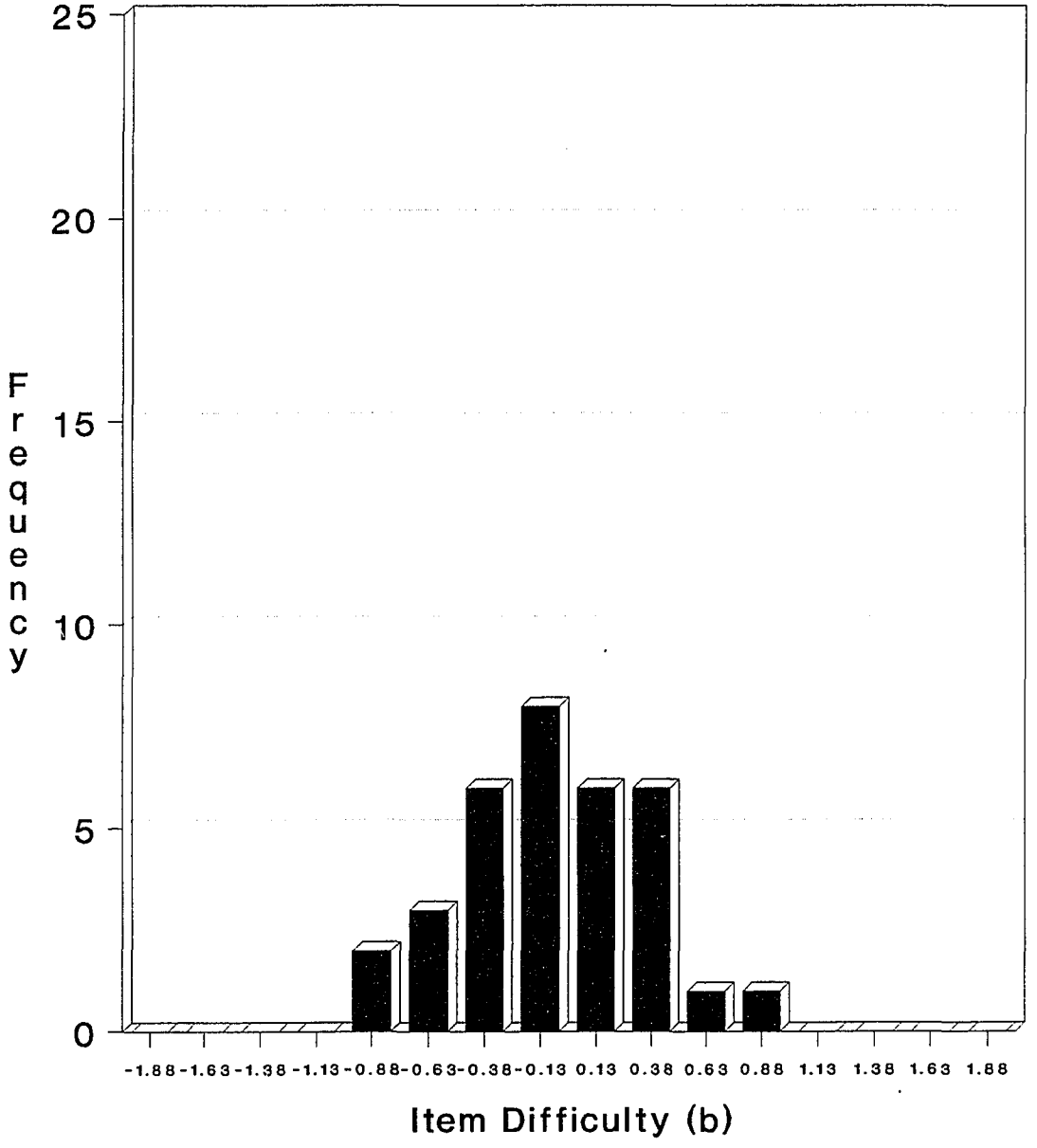


Figure 3
Item Difficulty Distribution of Calculations Examination B (33 Items)

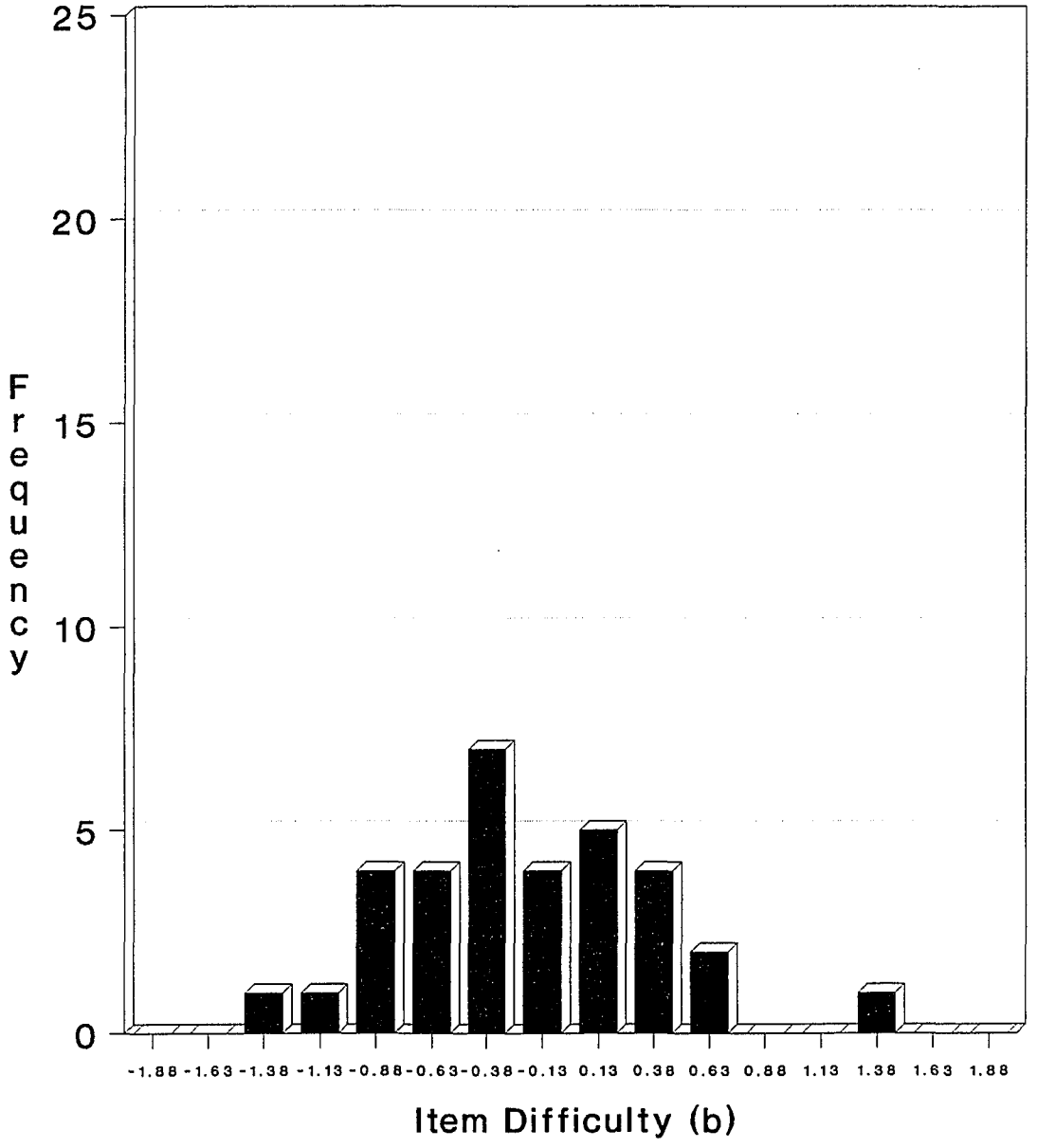


Figure 4
Item Difficulty Distribution of Calculations Examination C (35 Items)

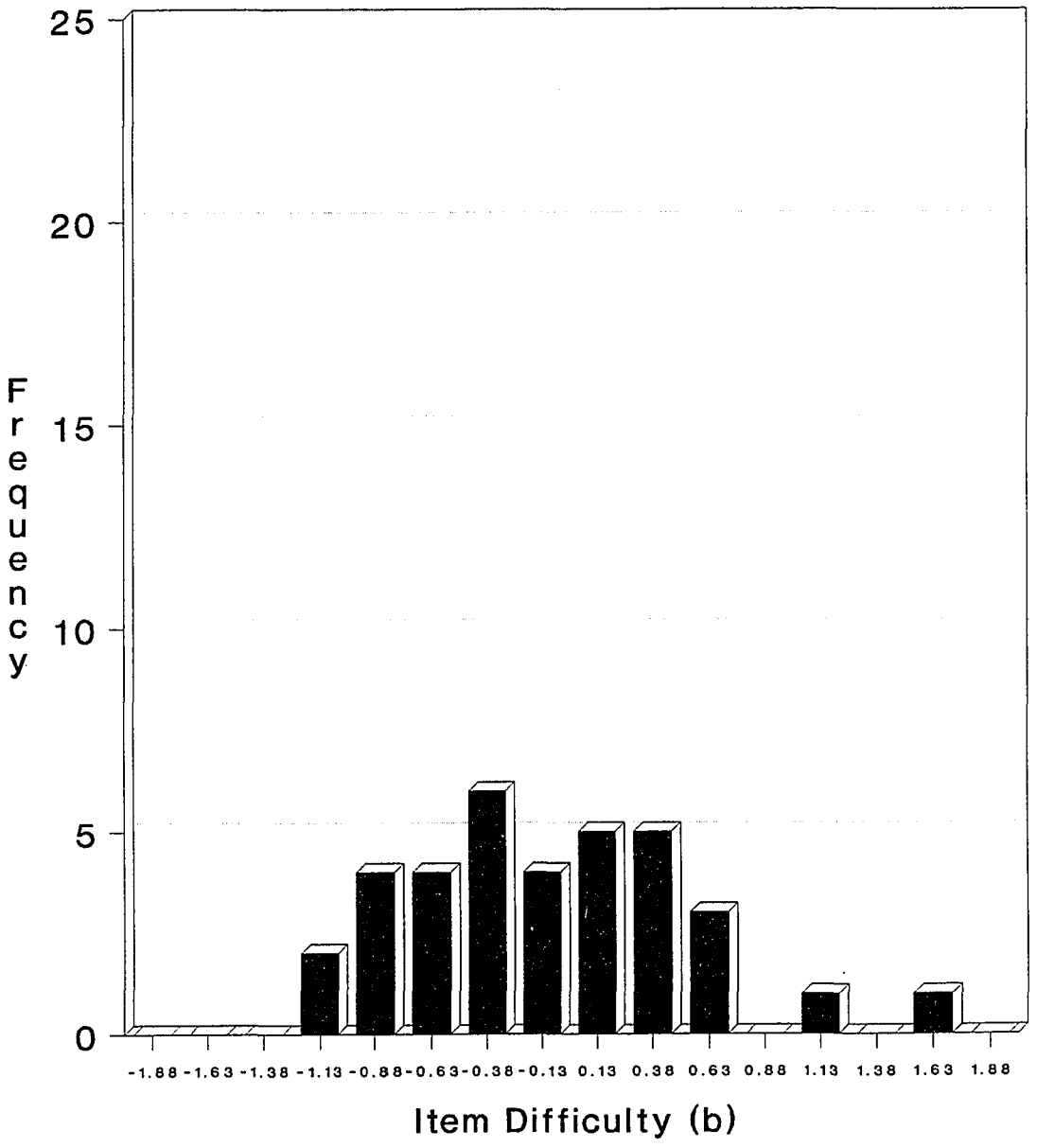


Figure 5
Item Difficulty Distribution of Calculations Examination D (36 Items)

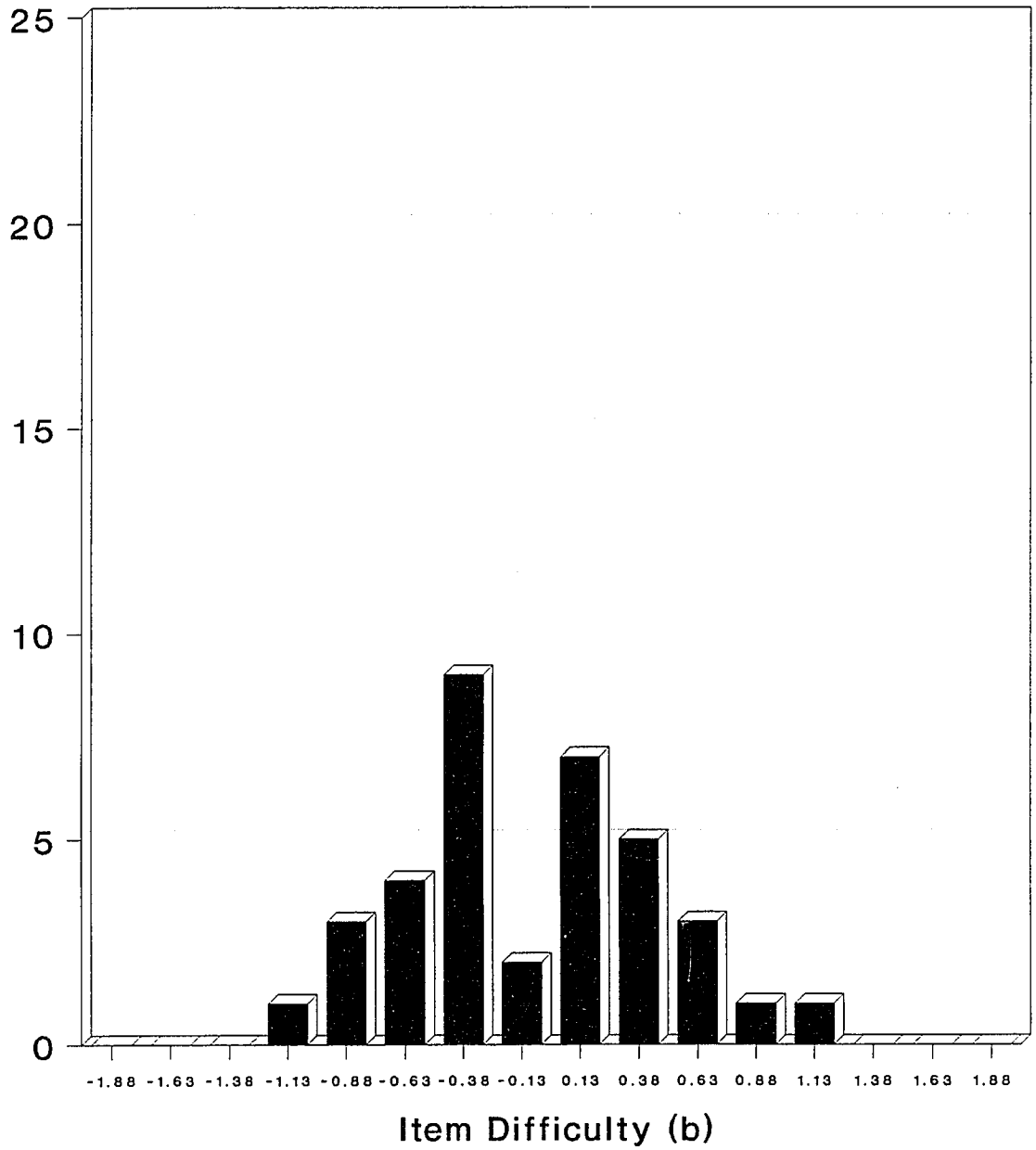


Figure 6
Information Function of Calculations Examination A (33 Items)

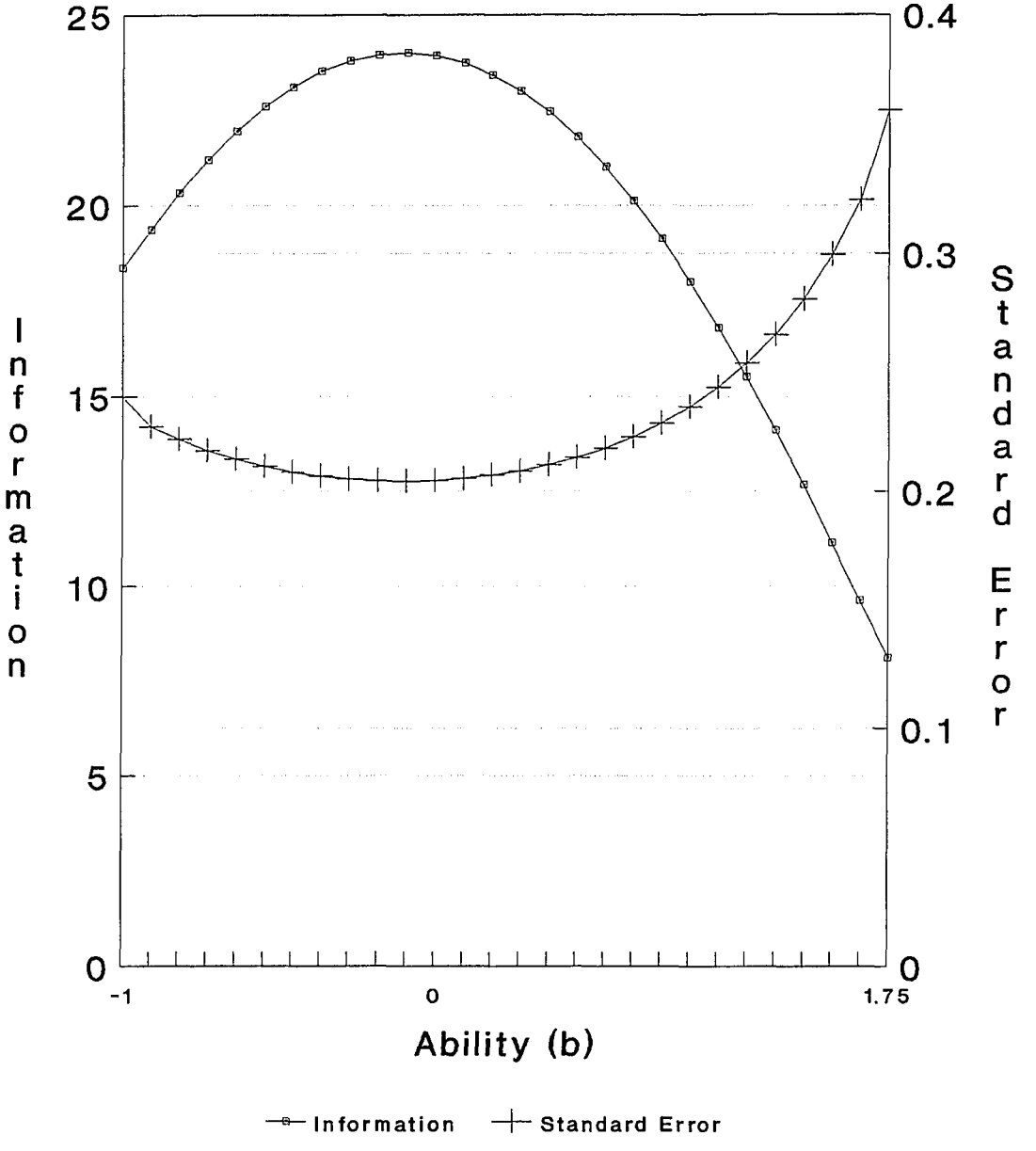


Figure 7
Information Function of Calculations Examination B (33 Items)

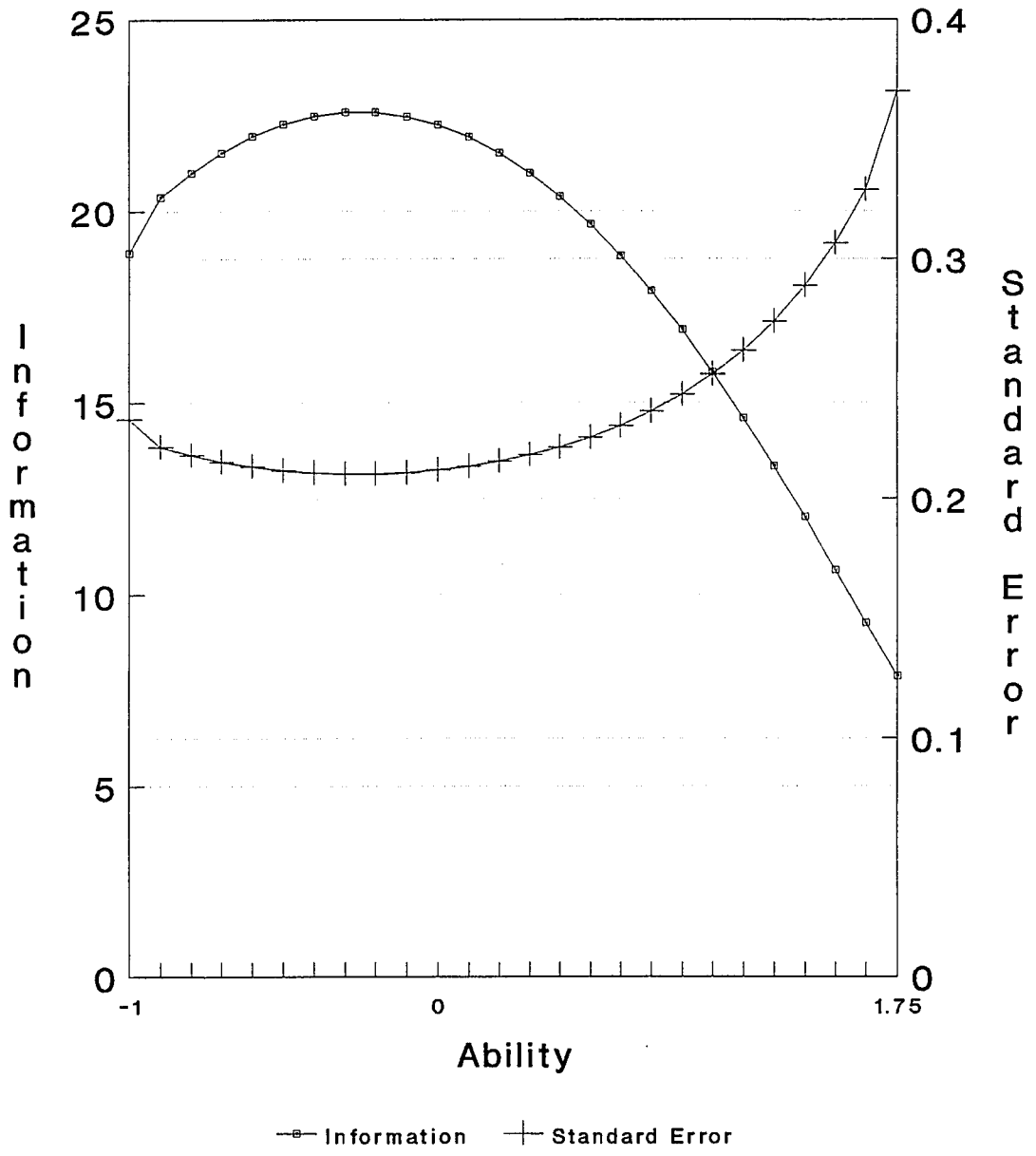


Figure 8
Information Function of Calculations Examination C (35 Items)

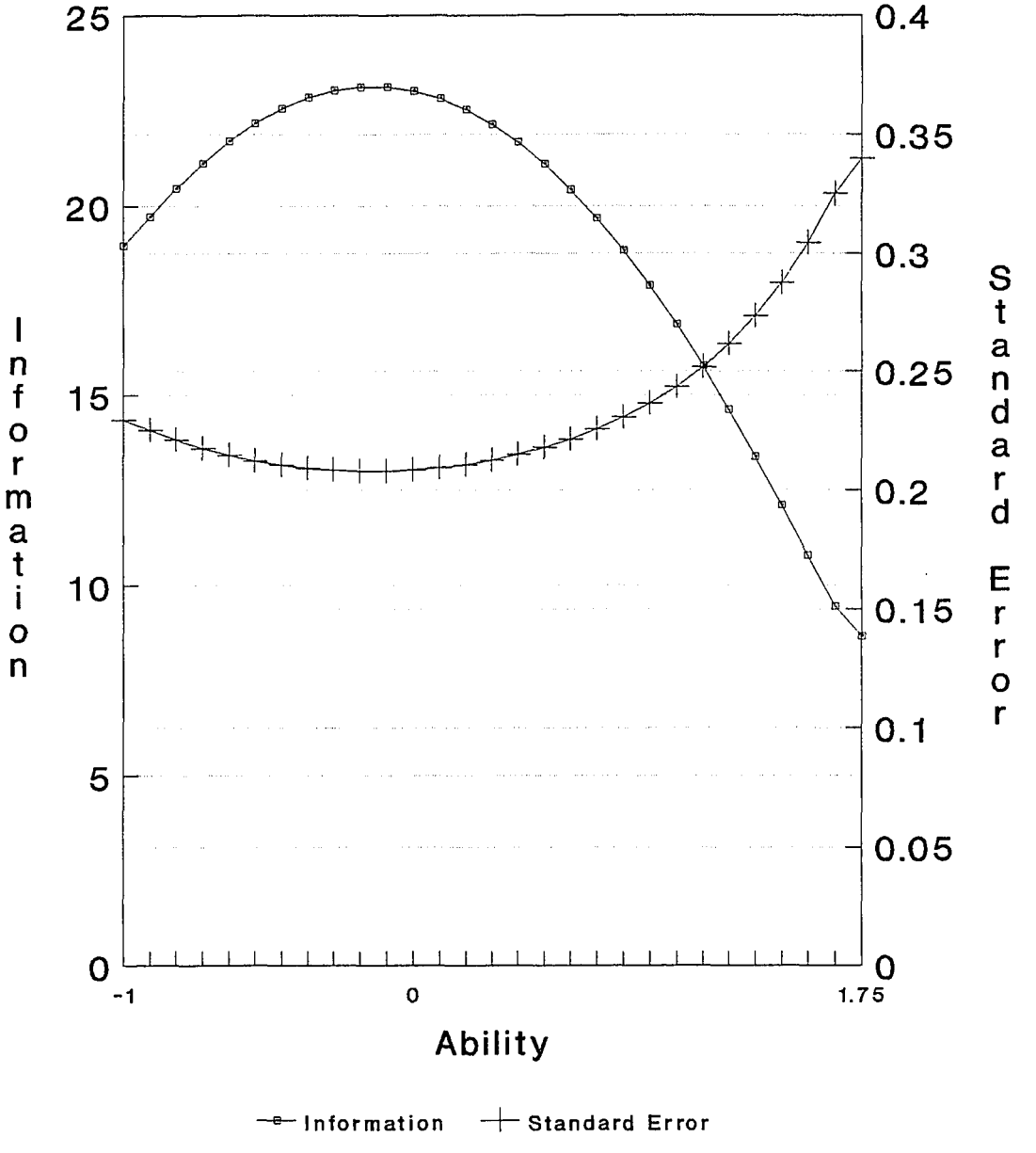


Figure 9
Information Function of Calculations Examination D (36 Items)

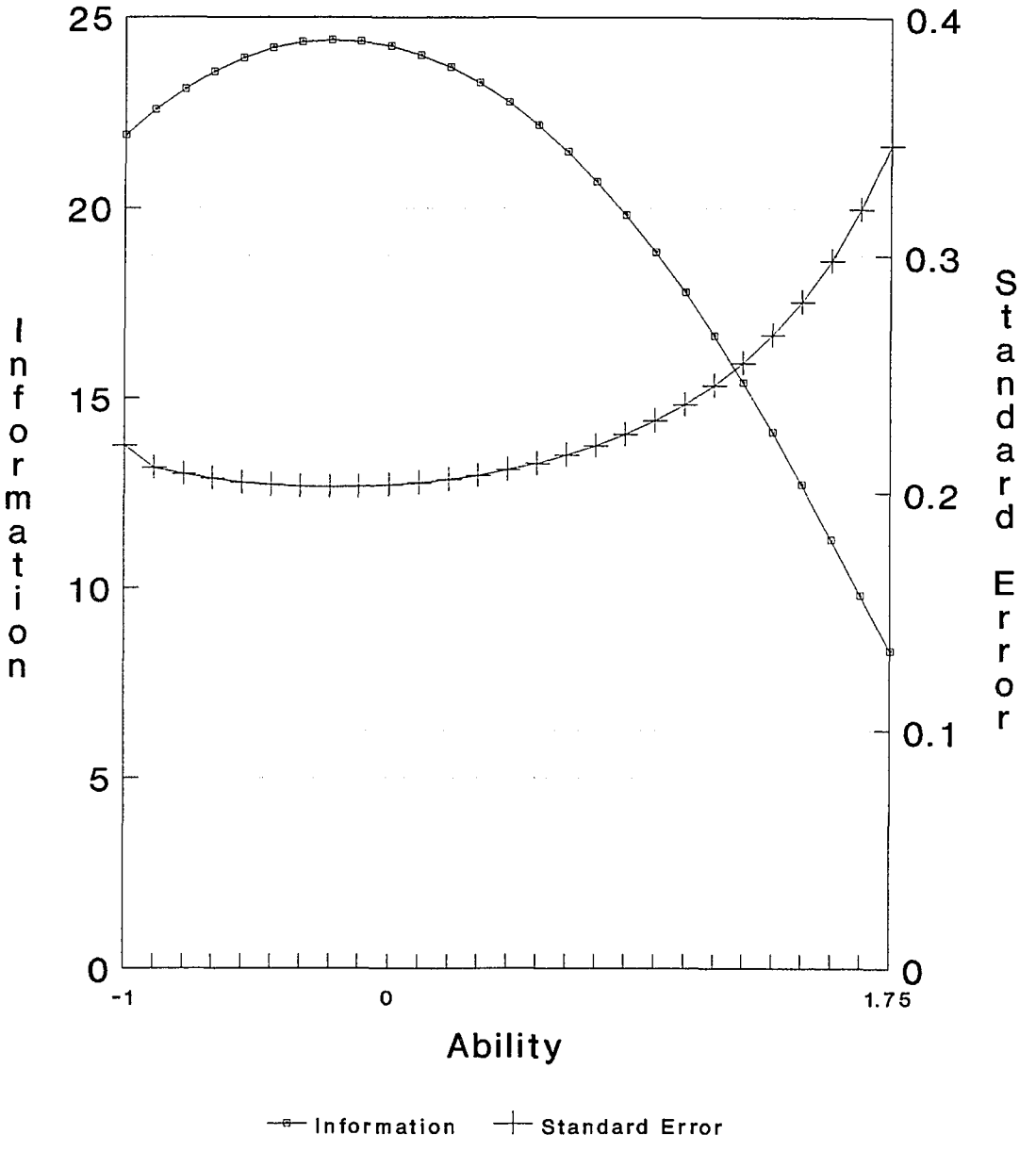


Table 4

Standard Errors of Estimate Achieved at Critical Points for Four Paper-and-Pencil Forms of Calculations of Drug Administration

Critical Point Ability	Form A	B	C	D
-1.00	.24	.23	.23	.23
-0.75	.23	.22	.22	.22
-0.50	.21	.21	.21	.21
-0.25	.21	.21	.21	.20
0.00	.20	.21	.21	.20
+0.25	.21	.22	.21	.21
+0.50	.22	.23	.23	.22
+0.75	.24	.25	.24	.23
+1.00	.27	.27	.26	.26
+1.25	.29	.30	.29	.28
+1.50	.33	.32	.31	.31
+1.75	.36	.37	.34	.35

Chapter IV

RESULTS

A direct comparison of CAT versus paper-and-pencil results was undertaken by comparing the SEMs achieved after the administration of *n* adaptive items to *n* length paper-and-pencil tests. Thus, the SEM at each critical point along the ability continuum was compared to the SEM ascertained at each of these ability levels at the end of the paper-and-pencil tests. Tables 5 through 7 provide a summary of the mean SEM at each of the critical points and the mean ability estimate for a 33, 35, and 36 item adaptive tests respectively, as well as the SEM for the paper-and-pencil administrations. A visual inspection of the data indicates that for every ability level as well as every test length the SEM achieved in the CAT simulation is smaller than that achieved in the paper-and-pencil administration.

The data in Table 5 indicate that on the 36 item (Form D) comparison, SEM ranges from .20 (b=-0.25, b=0.00) to .35 (b=1.75) on the paper-and-pencil examination. For the corresponding abilities on the CAT, SEMs range from a low of .192 to a high of .32. For each of the twelve critical points, the SEM ascertained via the CAT administration is lower than that achieved through paper-and-pencil administrations of the same item length. The smallest difference exists about the center of the ability continuum where the majority of the items are clustered.

In Table 6 the data demonstrate that on the 35 item comparison, the paper-and-pencil exam yields SEMs that range from a low of .21 for examinee abilities b=-.05, b-.25, b=0.0, and b=.25 to an SEM of .34 for an ability of b=1.75. The corresponding SEMS on the CAT are .20 and .32 respectively. Once again at each of

the twelve critical points, the mean SEM achieved through the CAT administrations are lower than those ascertained via paper-and-pencil administrations.

The 33 item exams are compared in Table 7 where a similar pattern is detected. On Form A, the SEM ranges from .20 to .36, and for the Form B SEM ranges from .21 to .37. On the CAT, SEMs corresponding to theses abilities range from .20 to .321. For each of the comparisons, both 33-items CATs yields smaller SEMs than the paper-and-pencil counterpart.

The gains in accuracy noted above are realized in Figures 10 through 13 which depict the information functions of the *n* length adaptive test along with their *n* length paper-and-pencil counterparts. Note that two such graphs are provided for the 33 item CAT as there are two 33 item paper-and-pencil tests to which to compare. In each instance, the information function of the CAT depicts greater accuracy along the entire ability continuum. In addition, the information function of the CATs are all flatter than those of the paper-and-pencil administrations, indicating that not only is the CAT more informative and accurate along the entire ability continuum but also that precision estimates about the entire continuum are more disparate in the paper-and-pencil administrations than the CAT. Information is maximized about the mean item difficulty for the paper-and-pencil tests and is less pronounced in ability regions where the number of items matching the ability is limited. This would confirm earlier notions that a CAT is more accurate in its estimation of ability and that this accuracy tends to be more equivalent between abilities on a CAT administration than on a conventional exam.

A numerical comparison of SEMs achieved in the conventional testing to those ascertained in CAT simulation is given in Table 8. The difference between the mean

Table 5

Mean SEM of 36-Item Adaptive Test and 36-Item Conventional Test (Form D) at Critical Points

Latent Ability	Paper-and-Pencil Standard Error Form D	Simulated CAT Standard Error (mean)	Simulated CAT Ability Estimate (mean)
-1.00	.23	.207	-1.031
-0.75	.22	.201	-.774
-0.50	.21	.197	-.499
-0.25	.20	.192	-.246
+0.00	.20	.198	-.130
+0.25	.21	.197	.254
+0.50	.22	.199	.506
+0.75	.23	.213	.757
+1.00	.26	.221	.989
+1.25	.28	.244	1.221
+1.50	.31	.274	1.525
+1.75	.35	.320	1.763

Table 6

Mean SEM of 35-Item Adaptive Test and 35-Item Conventional Test (Form C) at Critical Points

Latent Ability	Paper-and-Pencil Standard Error Form C	Simulated CAT Standard Error (mean)	Simulated CAT Ability Estimate (mean)
-1.00	.23	.210	-1.015
-0.75	.22	.202	-.799
-0.50	.21	.200	-.532
-0.25	.21	.200	-.228
+0.00	.21	.200	-.040
+0.25	.21	.200	.246
+0.50	.23	.203	.533
+0.75	.24	.214	.769
+1.00	.26	.222	.978
+1.25	.29	.245	1.219
+1.50	.31	.280	1.528
+1.75	.34	.320	1.755

Table 7

Mean SEM of 33-Item Adaptive Test and 33-Item Conventional Tests (Forms A & B) at Critical Points

Latent Ability	Paper-and-Pencil Standard Error Form		Simulated CAT Standard Error (mean)	Simulated CAT Ability Estimate (mean)
	A	B		
-1.00	.24	.23	.217	-1.021
-0.75	.23	.22	.208	-.786
-0.50	.21	.21	.206	-.504
-0.25	.21	.21	.200	-.242
+0.00	.20	.21	.204	0.000
+0.25	.21	.22	.201	.246
+0.50	.22	.23	.210	.593
+0.75	.24	.25	.222	.805
+1.00	.27	.27	.228	.956
+1.25	.29	.30	.248	1.204
+1.50	.33	.32	.284	1.510
+1.75	.36	.37	.321	1.736

Figure 10
Standard Errors of
CAT vs. Paper & Pencil A (33 Items)

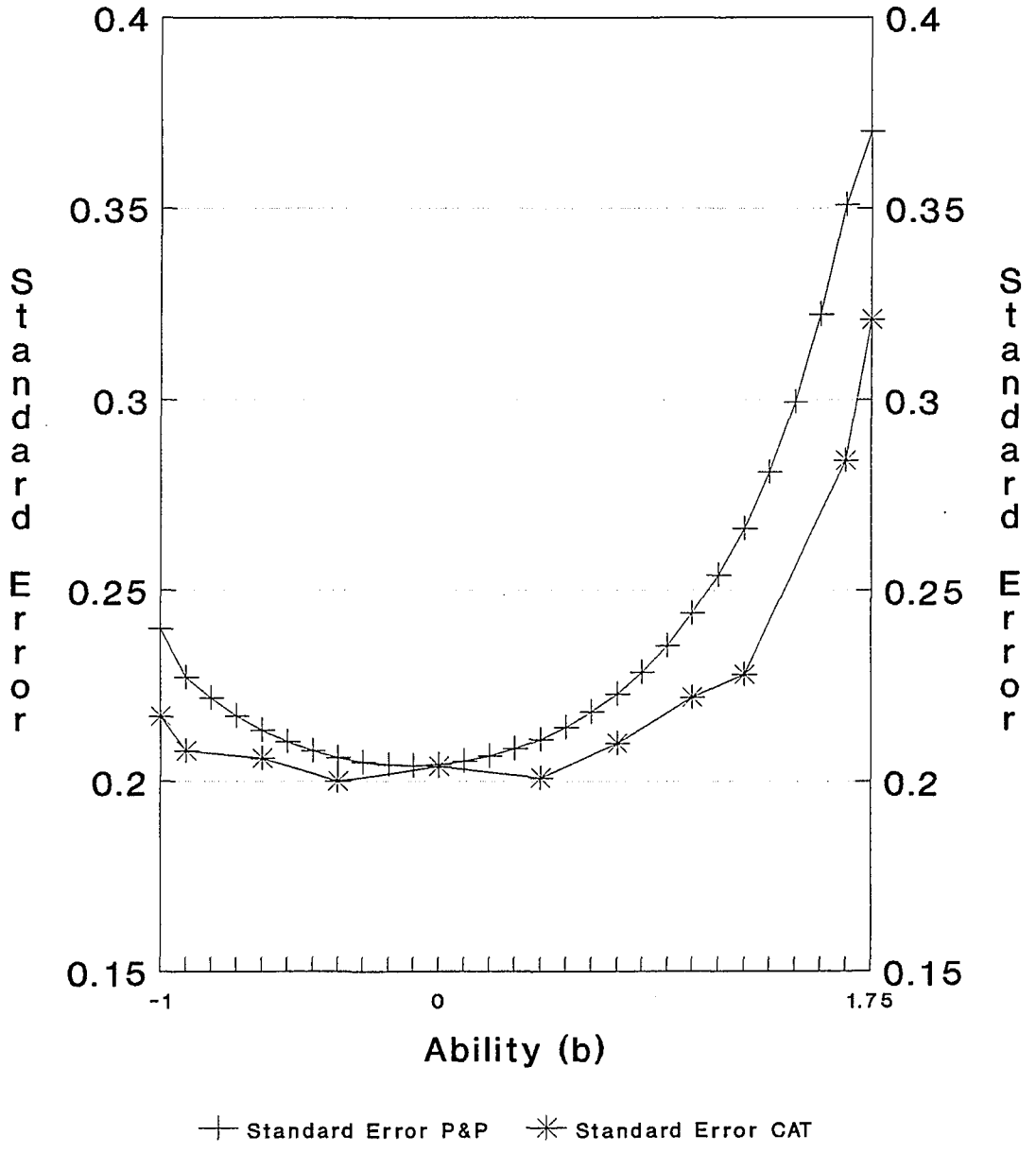


Figure 11
Standard Errors of
CAT vs. Paper & Pencil B (33 Items)

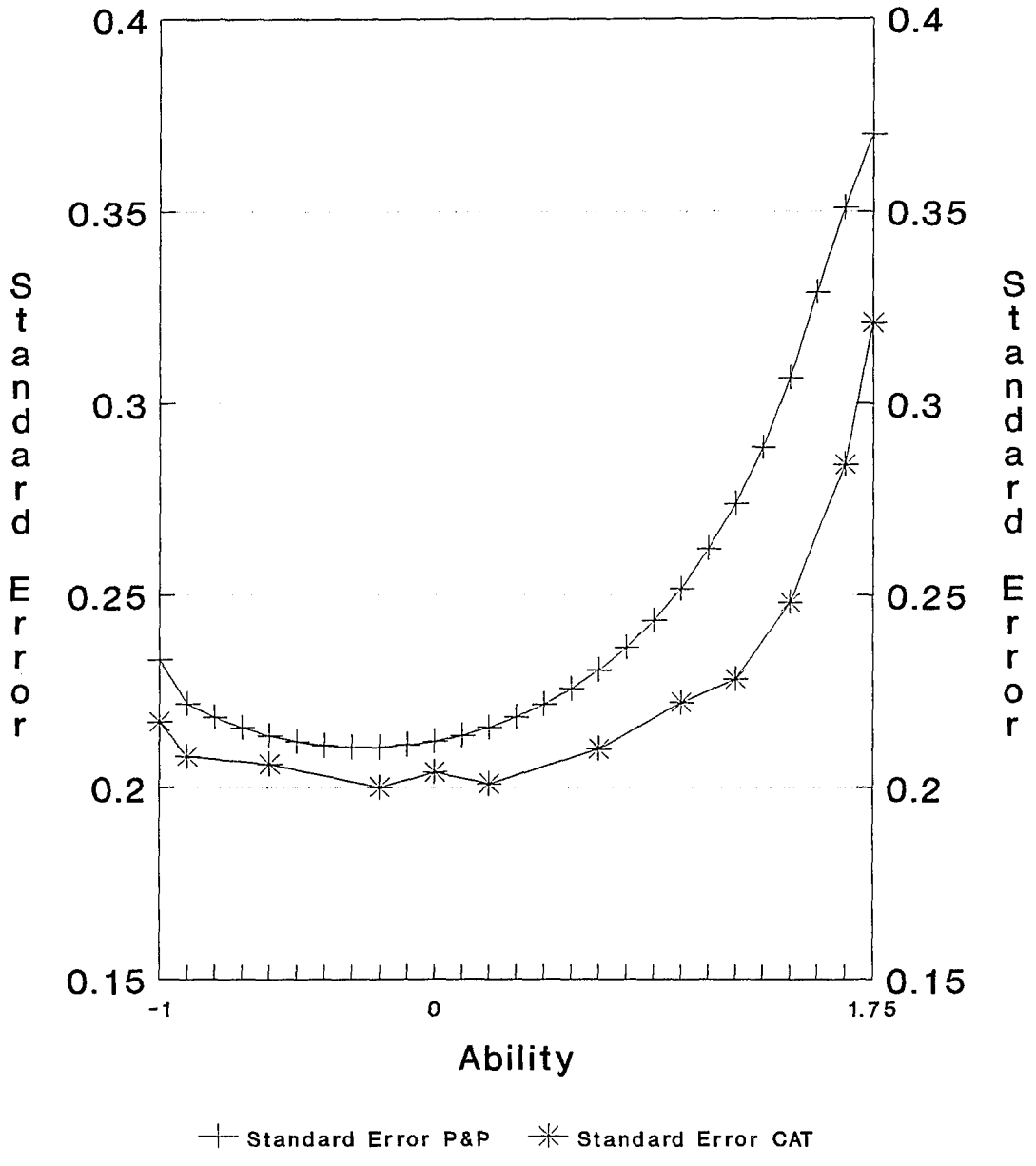


Figure 12
Standard Errors of
CAT vs. Paper & Pencil C (35 Items)

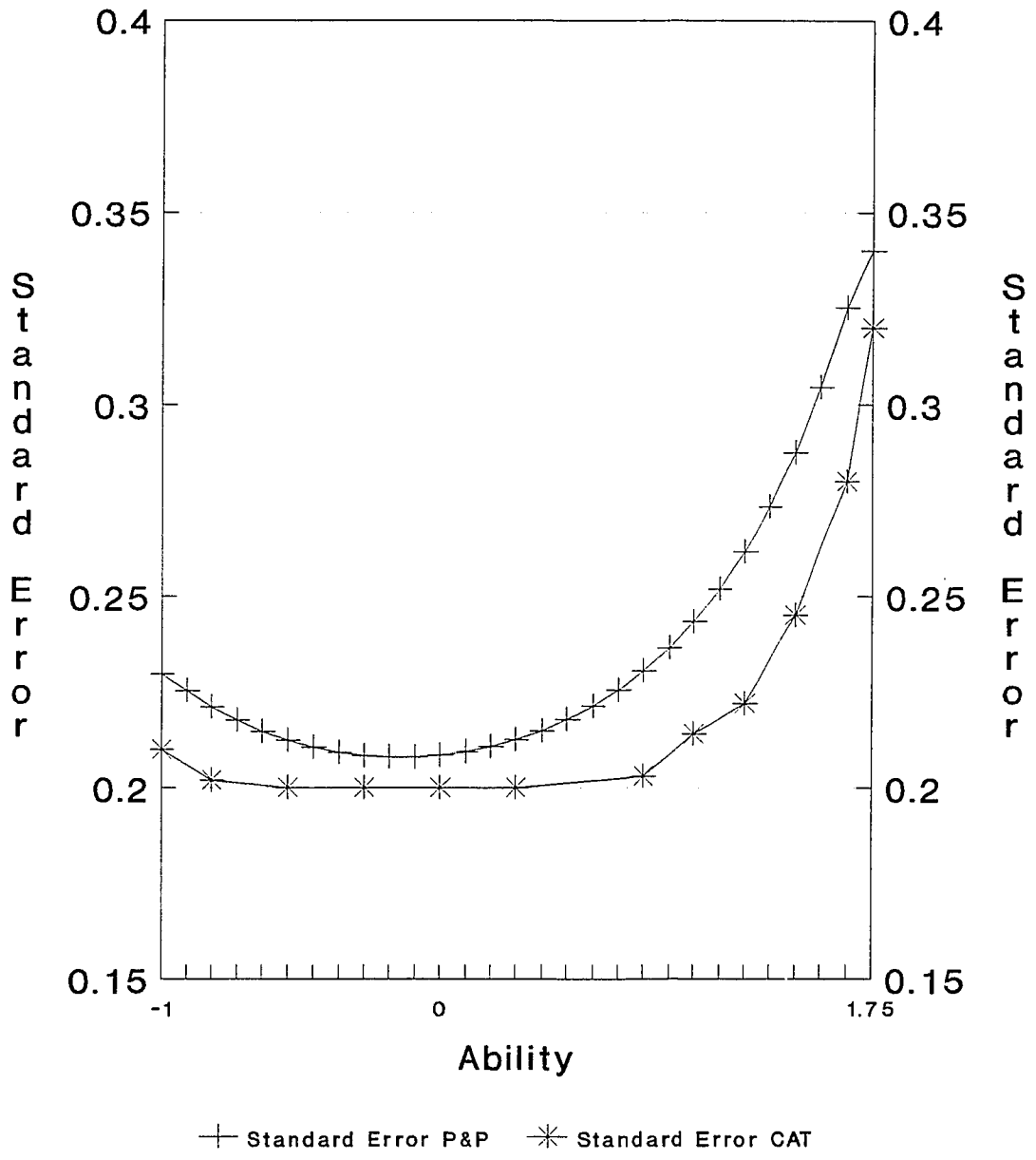
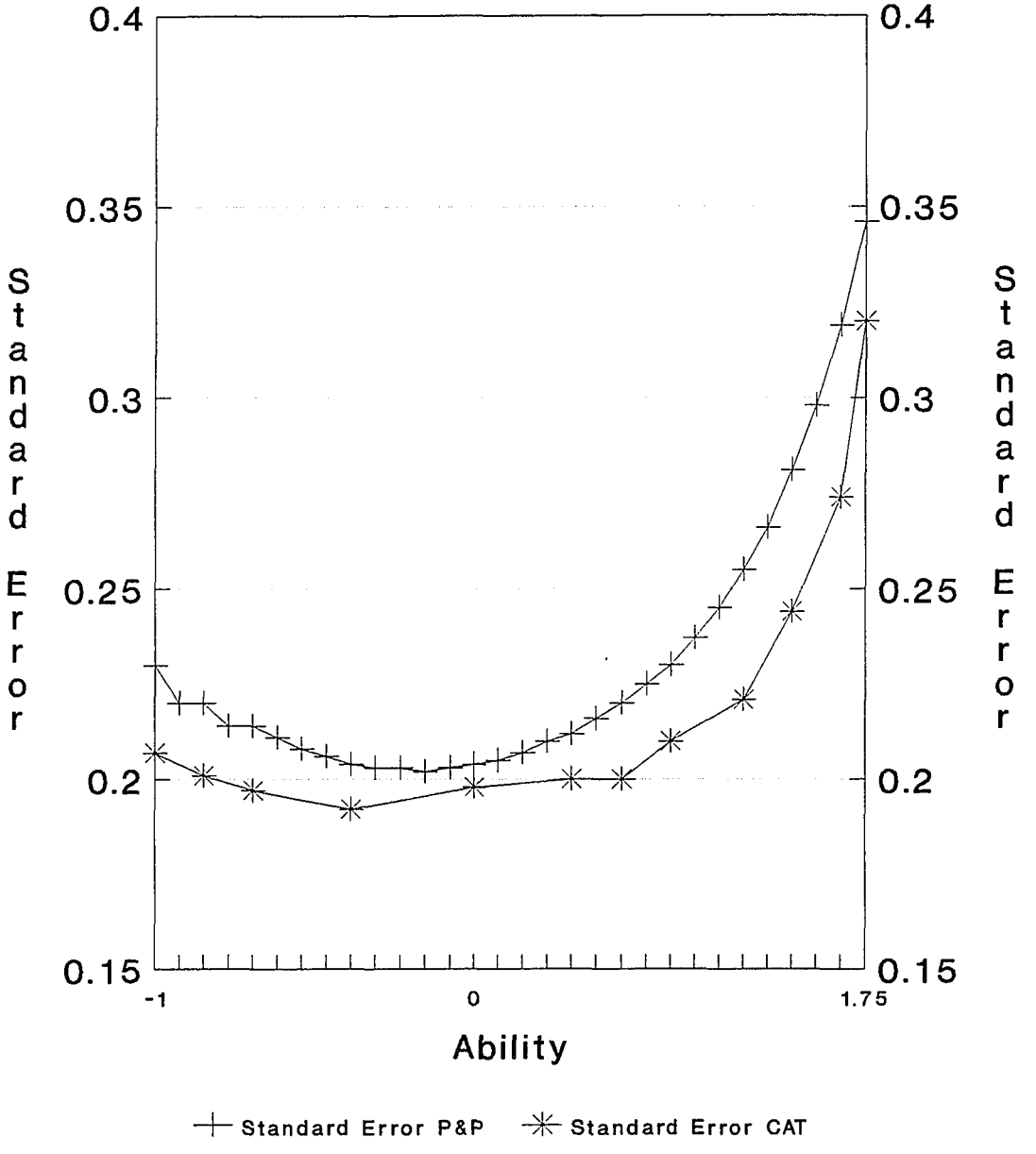


Figure 13
Standard Errors of
CAT vs. Paper & Pencil D (36 Items)



CAT SEM and mean paper-and-pencil SEM form the critical ability levels is provided in them first four columns for the 35 item, 36-item, and two 33 item exams. This is followed by four columns indicating a proportional comparison of CAT.SEM over paper-and-pencil SEM. For the 33 item examinations, the greatest difference in precision between the two testing frameworks is noted at the upper end of the ability continuum (1.25 to 1.75), where SEM is approximately 20% smaller on the CAT than on the paper-and-pencil exam. A similar pattern is noted for the 35 and 36 item exams. This phenomenon can be explained by considering the item difficulty distributions of the paper-and-pencil exams. On all four paper-and-pencil exams, items of higher difficulty level are underrepresented (See Figures 2-5). No items (0%) above 1.00 exist on Form A (33 items), where mean item difficulty is -0.063; on Form D (36 items), one item (2.7% of the item pool) above 1.00 exists, and mean item difficulty is -0.086; two items (5.7%) above 1.00 exist on Form C (35 items) where the mean item difficulty is -0.089; and finally on Form B (33 items), one item (3.0%) is above 1.00 and the mean item difficulty is -0.186. In addition, the difference in accuracy is less extreme about the center of the item difficulty distribution (approximately 0.00) where the majority of the items reside on each of the four paper-and-pencil test forms.

To consider the statistical probability of these results, the Sign test for two groups was undertaken for each of the four comparisons. The Sign test is used in which *N* pairs of matched observations are made to test the probability of a particular distribution of binomial possibilities (Hays, 1988). In this case, the question being put forth is whether the percentage of occasions that the CAT is superior in accuracy (SEM) is equal to the percentage of times that the paper-and-pencil exam is superior, in terms of SEM, along the ability continuum. In actuality, the null hypotheses being

Table 8

Comparison of Paper-and Pencil and CAT Standard Errors of Measurement for a Uniform Ability Distribution

Latent Ability	SEM DIFFERENCE (SEM_{CAT}-$SEM_{P\&P}$)				$(SEM_{CAT})^2/(SEM_{P\&P})^2$			
	33 Items		35 Items	36 Items	33 Items		35 Items	36 Items
	A	B	C	D	A	B	C	D
-1.00	-.023	-.013	-.020	-.023	.817	.890	.834	.810
-0.75	-.022	-.012	-.018	-.019	.817	.894	.843	.835
-0.50	-.004	-.004	-.010	-.013	.962	.962	.907	.880
-0.25	-.010	-.010	-.010	-.008	.907	.907	.907	.922
0.00	-.004	-.006	-.010	-.002	.961	.944	.907	.980
+0.25	-.009	-.019	-.010	-.013	.916	.835	.907	.880
+0.50	-.010	-.020	-.027	-.012	.911	.834	.779	.818
+0.75	-.018	-.028	-.026	-.017	.856	.789	.795	.858
+1.00	-.042	-.042	-.038	-.039	.713	.713	.729	.723
+1.25	-.042	-.052	-.045	-.036	.731	.713	.714	.759
+1.50	-.046	-.036	-.030	-.036	.741	.788	.816	.781
+1.75	-.039	-.049	-.042	-.030	.795	.753	.886	.836

tested is the probability that CAT is superior to paper-and-pencil exam in terms of accuracy estimates is equal to 50% (chance).

In fact, for each of the four comparisons, the CAT SEMs are superior to those of the paper-and-pencil exam at each of the twelve critical points along the ability continuum ($p = 1.00$). With twelve matched pairs, the probability of this occurring is 0.0002 and thus the null hypothesis is rejected at both the alpha = .05 and alpha = .001 level, suggesting that each CAT is superior to each paper-and-pencil administration regardless of the number items administered (33, 35, or 36).

Discussion for the results thus far has been based on the assumption of a uniform distribution of ability. Results also were considered in light of non-uniform distributions to determine the advantage of CAT with limited item pools in those situations. To undertake these analyses, five non-uniform, normal distributions were considered (0,1), (.5,1) (1,1) (-.5,1) (-1,1). Precision estimate (SEM) gains of the CAT over each of the four paper-and-pencil administrations from the initial simulation were weighted to determine a net gain for each distribution.

Mean gains in SEM are provided when we assume both uniform and non-uniform simulee ability distributions. For the uniform distribution, the net gain is simply the mean gain for each of the twelve critical points; these gains are then weighted appropriately for each of the five non-uniform distributions. Mean gains are provided for both 33-item paper-and-pencil exams with the 33 item CAT, 35-item paper-and-pencil exam with the 35-item CAT, and 36-item paper-and-pencil exam with the 36-item CAT.

Table 9 provides the mean gains in SEM when considering the uniform and four non-uniform distributions. When CAT measures of precision are compared to these

indices on the paper-and-pencil exams, the greatest gains are noted when we assume the distribution of ability to be Normal (1,1), i.e. high ability examinees. A mean gain of .040 SEMs is achieved when the 33-item CAT is compared to the 33-item paper-and-pencil exam (A); the gain is .040 SEMs when the CAT is compared to version B (33 items), .045 SEMs when compared to the 35-item paper-and-pencil exam, and 0.37 SEMs when compared to the 36-item paper-and pencil exam. The smallest gains in accuracy are noted when the CAT measures of precision are compared to those indices on a paper-and-pencil exam, when we assume the distribution of simulee ability to be Normal (-1,1), i.e low ability examinees. The gain when the 33-item CAT is compared to version A (33-item paper-and-pencil exam) the gain is .016 SEMS. The mean gain is .015 SEMs when the other 33-item exam (B) is compared to the precision indices of the CAT, .018 for the 35 item comparison, and .018 for the 36-item comparison.

These results suggest once again that the greatest gain of the CAT over the paper-and-pencil administrations is noted at the upper level of the ability continuum and can be attributed to the fact that limited items exist in this region on each of the four paper-and-pencil forms. Thus, larger SEMs are achieved at the upper ability continuum. However, when all items are combined into a larger CAT pool, not only does the number of items available for selection increase, but also the adaptive aspect of the CAT allows us to selects items that are more matched to these upper region abilities, minimizing the number of less informative items administered to the examinees, and therefore increasing the accuracy of ability estimates for these examinees. This notion is further confirmed by the fact that the N(.5,1) ability distribution demonstrates the greatest gain in accuracy after the N(1,1) distribution. Mean gains for the former range

from .035 when the 36-item exams are compared to .044 when 33-item paper-and-pencil exam (Form B) is compared to the 33 item CAT. Not surprisingly, the N(-.5, 1) and N(-1,1) distributions demonstrate the least amount of accuracy enhancement when the paper-and-pencil exams are compared to the CAT. For the N(-5,1) distribution mean gains range from .015 to .018 and for the (-1,1) distributions gains range from .022 to .026. These gains in accuracy of the CAT over the paper-and-pencil exams are even less than those attained when we assume a uniform distribution, and suggests again that we note less enhancement in CAT when the paper-and-pencil exams consist of items that range in a difficulty continuum that more closely matches the ability distribution of the examinees who are being tested. The composition of all for paper-and-pencil exams demonstrate this characteristic. Yet even with this situation, CAT still provides a slight enhancement in accuracy.

In the second analysis of the results, the data were examined by comparing the number of adaptive items need to achieve the same precision as each *n*-length paper-and-pencil tests. Thus, the data were analyzed by determining the mean SEM after the administration of *n* traditional items and then determining at which point along the adaptive simulation each examinee ascertained this SEM. The mean number of CAT items required to achieve the same precision as each of the paper-and-pencil exams at each of the critical points are listed in Table 10.

When compared to paper-and-pencil exam version A (33 items), the CAT requires a smaller number of items to achieve the same level of precision as the traditional administration. Only at a latent ability of 0.00 is the number of items the same. Overall, the CAT requires a mean number of 25.6 items as compared to the 33 items of the paper-and-pencil exam. Again, the results are most dramatic at the extreme

Table 9

Mean Gain in Precision (SEM) Using CAT for Theoretical Uniform and Non-Uniform Ability Distributions

	Theoretical Distributions					
	Uniform	Normal (-1,1)	Normal (-.5,1)	Normal (0,1)	Normal (.5,1)	Normal (1,1)
33-item CAT (compared to A)	.022	.016	.022	.031	.036	.040
33-item CAT (compared to B)	.026	.015	.023	.034	.044	.050
35-item CAT	.025	.018	.026	.032	.042	.045
36-item CAT	.022	.018	.024	.030	.035	.037

ability levels, where the number of available items is limited. At an ability of 1.75 the 33 item paper-and-pencil exam (Form A) achieves a precision of .36 SEM; this level of precision is achieved with a CAT of 18.6 items. A similar result, although less extreme, is evidenced also at ability level -1.00, where only 25.2 items are required to achieve the precision of the 33 item paper-and-pencil exam. In the middle of the ability/difficulty continuum where there a larger number of items on the paper-and-pencil exams, the number of CAT items required to achieve equiprecision is also fewer but less dramatic; the tailoring aspect of the CAT has less impact, yet even here fewer CAT items are required.

Similar results are achieved when the CAT administration is compared to the other 33-item paper-and-pencil administration (Form B), 36-item paper-and-pencil exam (Form D), and the 35-item traditional administration (Form C). When one considers the overall mean number of CAT items needed to achieve equiprecision about the entire ability continuum, eight to ten less items are required to achieve the same level of precision as the paper-and-pencil administrations. Approximately 25 items are required by the CAT to achieve the same level of precision of the 33 item traditional examinations; approximately 26 items are need to achieve the same precision level as the 35 item paper-and-pencil exam; and approximately 27 items are need by the CAT to achieve the same level of precision as the 36-item paper-and-pencil exam. This confirms earlier notions that CAT achieves equal levels of precision as traditional paper-and-pencil tests with a smaller number of items administered.

Figures 14 through 17 depict comparisons of the number of CAT items required to achieve the precision of the fix length, non-adaptive tests. As is suggested by the previously noted data, fewer adaptive items are required to achieve the precision of the

non-adaptive tests along the entire continuum. All four figures demonstrate that fewer adaptive items are needed to achieve the SEMs achieved on the non-adaptive exams. As was the case in the previous analyses, the greatest reduction in required items is noted at the upper end of the ability continuum. This fact corroborates results of the first analysis in which the greatest reduction in SEM is noted at the upper end of the ability continuum. Towards the center of the continuum, the number of adaptive items needed to achieve the precision of the non-adaptive test is larger. This fact, too, corroborates the fact that the smallest reduction in SEM when using the adaptive procedure also is noted about the center of the ability continuum.

Table 10

Mean Number of Items Required to Achieve Equiprecision with CAT as Compared to Paper-and-Pencil Administrations

Latent Ability	SEM Form A(33)	Items Needed by CAT	SEM Form B(33)	Items Needed by CAT	SEM Form C(35)	Items Needed by CAT	SEM Form D(36)	Items Needed by CAT
-1.00	.24	25.2	.23	27.4	.23	27.4	.23	27.4
-0.75	.23	26.6	.22	28.8	.22	28.8	.22	28.8
-0.50	.21	30.8	.21	30.8	.21	30.8	.21	30.8
-0.25	.21	29.3	.21	29.3	.21	29.3	.20	32.4
0.00	.20	33.2	.21	30.2	.21	30.2	.20	33.2
+0.25	.21	29.8	.22	27.2	.21	29.8	.21	29.8
+0.50	.22	28.6	.23	25.6	.23	25.6	.22	28.6
+0.75	.24	26.2	.25	24.7	.24	26.2	.23	28.3
+1.00	.27	19.6	.27	19.6	.26	21.3	.26	21.3
+1.25	.29	18.9	.30	17.1	.29	18.8	.28	19.2
+1.50	.33	19.6	.32	20.0	.31	22.0	.31	22.0
+1.75	.36	18.6	.37	15.7	.34	20.1	.35	19.1
MEAN		25.5		24.7		25.7		26.7

Figure 14
CAT Items Needed to Achieve Precision of Examination A (33 Items)

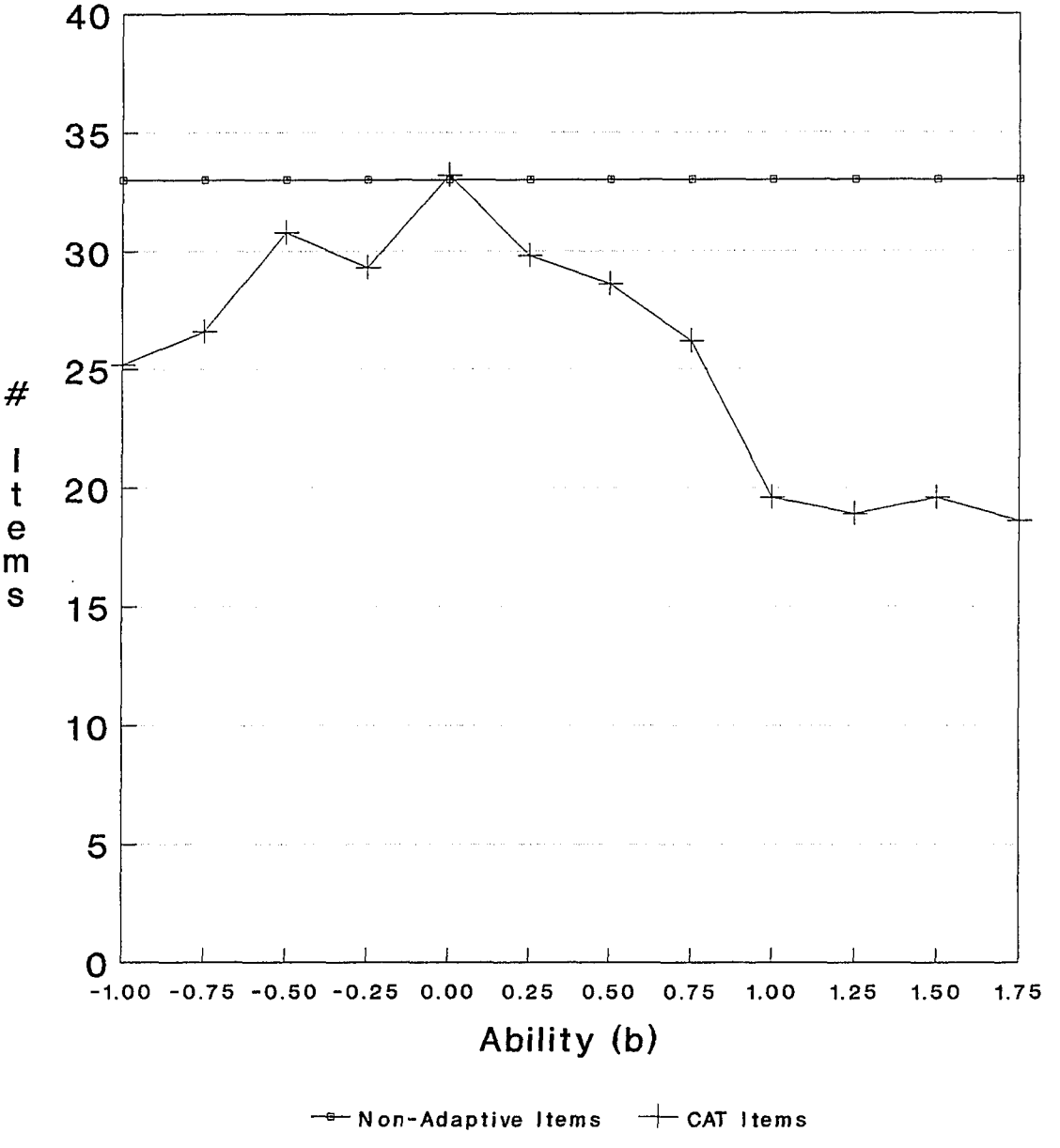


Figure 15
CAT Items Needed to Achieve Precision of Examination B (33 Items)

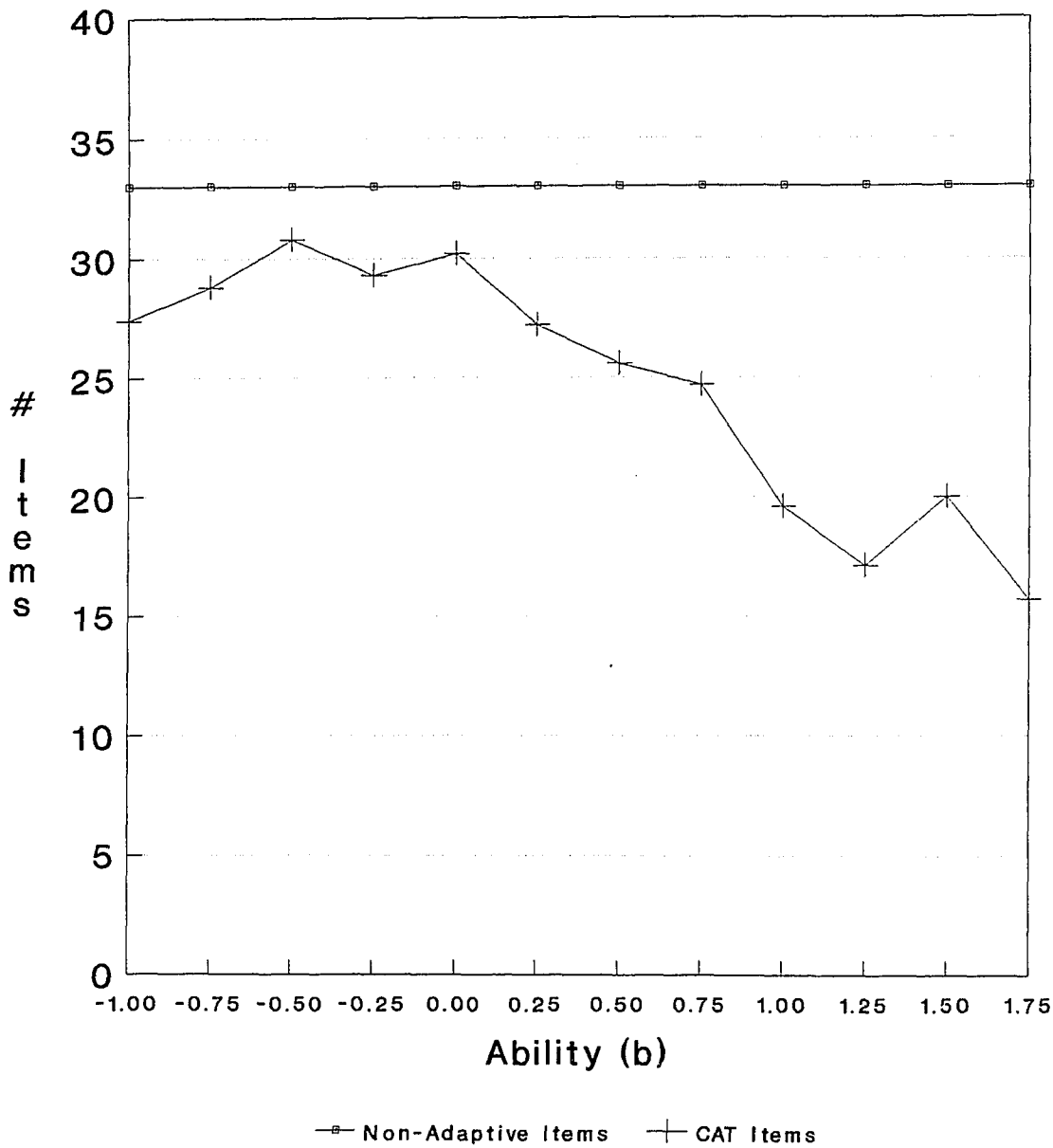


Figure 16
Items Needed to Achieve Precision of Examination C (35 Items)

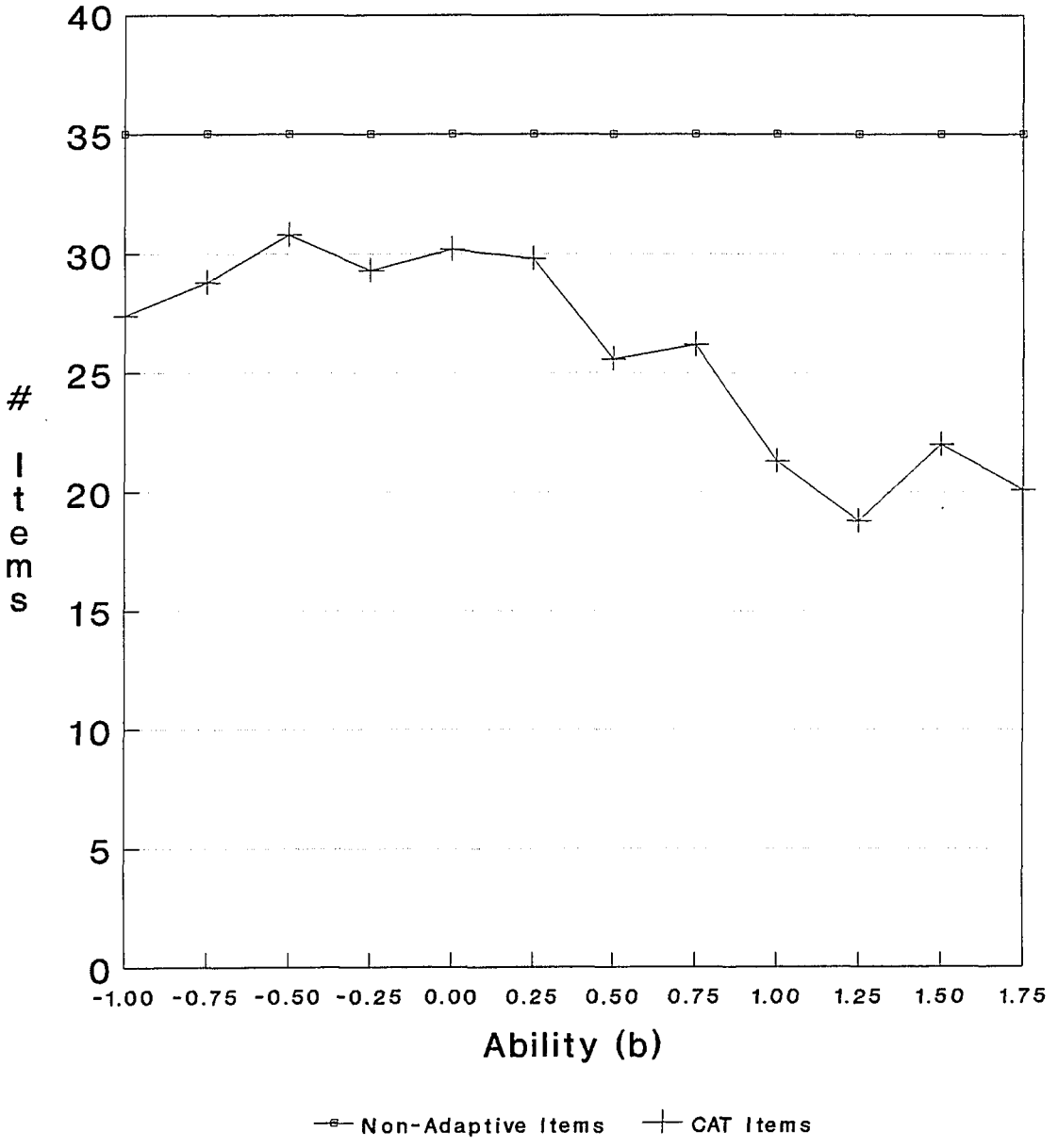
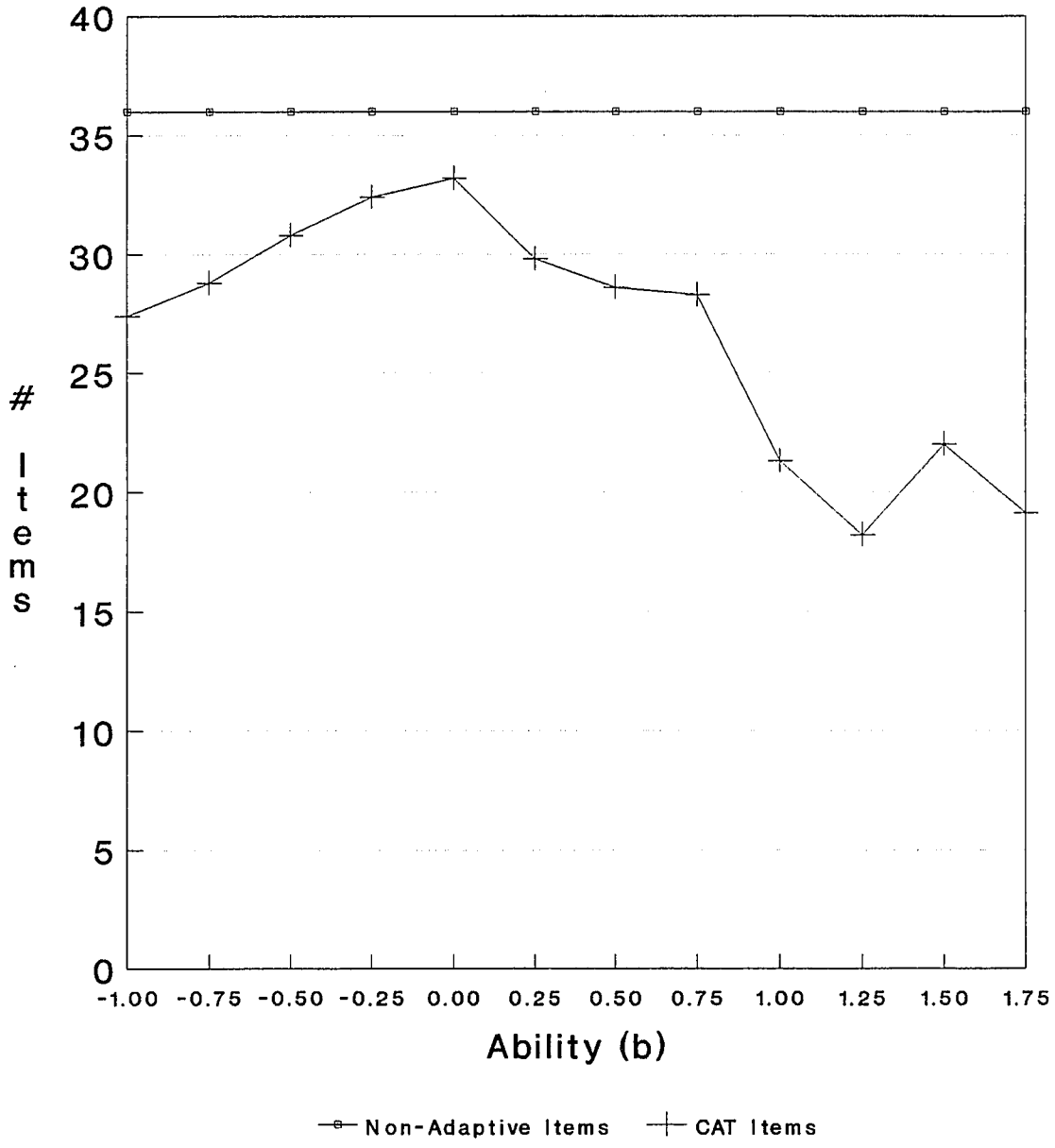


Figure 17
CAT Items Needed to Achieve Precision of Examination D (36 Items)



Chapter V

SUMMARY AND DISCUSSION

In order to determine the effectiveness of adaptive testing in situations where item pools are limited, a simulation study was conducted comparing the effectiveness of the CAT with a pool of 101 items to four paper-and-pencil exams consisting of 33, 35, or 36 items. The ability estimates of simulees, with presumed latent abilities ranging from -1.00 to +1.75, were generated along with the standard errors of measurement after the administration of 33, 35, and 36 tailored items. The items were drawn from the 101 without replacement. These accuracy estimates were compared to those attained in a live testing situation of 4371 examinees who took one of the four versions of the paper-and-pencil examination.

Comparisons of the two testing frameworks were undertaken by considering the SEMs for the paper-and-pencil exams as compared to those generated in the adaptive framework. Thus, each *n*-length paper-and-pencil exam was compared to the *n*-length adaptive exam in terms of SEMs achieved by the end of the testing process. In addition, the number of adaptive items required to achieve the accuracy level of the paper-and-pencil exams was determined. Initial comparisons were undertaken assuming an a uniform ability distribution ranging from -1.00 to +1.75 logits.

In both of the above sets of analyses, the CAT proved to be superior to the paper-and-pencil administration both in terms of accuracy at the end of the administration and in terms of the quantity of items required for the CAT to reach the level of accuracy of the tradition administrations. The mean level of accuracy for the 33-item paper-and-pencil examinations were .251 and .255 SEMs as compared to a

mean SEM of .229 achieved by the adaptive testing; for the 35-item administrations, mean SEMs for the traditional and adaptive exams along the examined ability continuum were .248 and .225 SEMs respectively; the 36-item CAT ascertained a mean SEM of .222 while the paper-and-pencil exam achieved an mean SEM of .244 after the administration of all 36 items. Thus, the adaptive exams achieve greater accuracy than the traditional exams; and, as would be expected, in both types of administrations the accuracy was enhanced as the item length increased, explaining the slight enhancement in accuracy of the 36-item examinations in both testing frameworks.

In terms of the number of items required to achieve equiprecision at each of the examined ability levels, an average of 25.5 adaptive items is required to achieve the precision of the 33-item tradition exam A, 25.7 adaptive items are required to achieve the precision of 33-item exam B and 35-item exam C, and an average of 26.7 items is required to achieve the precision of 36-item exam D. In each of these comparisons, approximately eight to ten fewer items are required by the adaptive exams to achieve the level of accuracy of the traditional exams.

The enhanced accuracy of the adaptive examinations at each of the ability levels over the paper-and-pencil examinations was confirmed statistically through the non-parametric S ign test and was also evidenced when the distributions of ability were assumed to be non-uniform in nature.

Similar results were suggested graphically by comparing the information function of each of the *n*-length paper-and-pencil exams with their adaptive counterparts. In all cases, the SEM curve for the adaptive exam results are lower and flatter than those attained with the paper-and-pencil administrations. However, the

difference in accuracy estimates is greater at the extremes of the ability continuum (approaching -1.00 and +1.75). In the middle of the ability continuum, (-.50 to +.50) is less dramatic. These results should be expected given that the distribution of item difficulties for each of the non-adaptive exams is centered about this range. Thus, the tradition examinations, by the very nature of the item distributions, may provide much information for achieving high accuracy estimates. By pooling the items into one CAT pool, the net result is less dramatic and results in only slightly more accurate estimates due to the tailoring. In fact, the enhancement of accuracy within this region ranges only from a minimum reduction of .004 logits to a maximum of .019 logits. The enhancement is greater as one moves away from this ability range to the extreme ability levels examined thus accounting for an overall gain in accuracy by adapting the examination. This idea is supported when one assumes the ability distribution to be non-uniform. The greatest gain in accuracy is noted when one assumes an ability distribution N(1,1). Thus when one assumes most examinees' abilities reside at the upper extreme, the overall gain in accuracy is greater because it is at this point in the ability continuum that the difference in SEM is greatest.

While the overall comparisons of the adaptive testing to the non-adaptive testing suggest greater precision in measurement by the adaptive exam, the results also depict an inequality of precisions about the entire ability continuum. Like the paper-and-pencil examinations, which lacked items at the high end of the ability continuum resulting in higher SEMs at these abilities, so too the adaptive examination failed to achieve equiprecision about the entire continuum with much lower measures of accuracy being ascertained at levels greater than +1.25. While the SEM function curve is lower than that of its paper-and-pencil counterparts at these abilities, the adaptive exam still fails

to achieve a precision comparable to those achieved at the lower ability levels. Thus it might be suggested that while the adaptive examination, even with its limited pool, does enhance accuracy when compared to non-adaptive administration, it is not effective in achieving equiprecision about the entire ability continuum when the item pool is limited in its composition.

While the above results would clearly establish the superiority of adaptive testing to paper-and-pencil testing, where all examinees takes the same set of items, it does nonetheless demonstrate the limitations of the approach. Like all tests that are as good as the items that constitute them, the effectiveness of CAT is dependent upon the item pool from which questions can be drawn. And while the adaptive test by its very nature of tailoring is sure to yield smaller SEMs, the framework in and of itself cannot provide an ultimate solution for accuracy of measurement at ability levels where these is a lack of items of equivalent difficulty. This result is clearly demonstrated by this study at the upper end of the ability continuum where the CAT provides greater accuracy than the paper-and-pencil exam, but still fails to achieve equiprecision to lower ability levels. It might, in fact, be argued that the sole reason that the CAT does provide a greater level of accuracy at these extreme abilities is due simply to the fact that all 101 items are available for administration. In other words, the most difficult items from all four paper-and-pencil exams are available for administrations. Conversely, on average only one-quarter of these difficult items are available on each of the four paper-and-pencil exams, thus the increased accuracy of the CAT at these levels is not surprising. In fact, this is the very reason that increased accuracy is achieved at all of the examined ability levels, and the reason that the comparison that assumes an ability distribution N(1,1) demonstrates the greatest advantage of the

adaptive over the non-adaptive testing.

One might argue that this is one of the basic tenets of adaptive testing--to pool all available items and select from among them as needed for each examinee, as compared to selecting *n* items to create an *n*-length paper-and-pencil test. The true test of the adaptive framework in its ability to overcome the limitations of a restricted pool, however, is not noted. The CAT parallels the results of the traditional tests in the shape of the information function, but simply increases its height due to the fact that the items are tailored and a larger selection of items is available for administration. These results corroborate previous findings (Stocking, 1987) where 20-item adaptive exams were compared to 20-item non-adaptive exams utilizing a variety of item difficulty distributions.

Nonetheless, the adaptive testing framework does enhance accuracy along the entire ability continuum, albeit the limited nature of the item pool prevents the theoretical equiprecision about the entire continuum to be achieved. Even when there are only 101 items from among which to choose, adapting an examination via mechanisms such as CAT will allow for greater precision with a smaller number of items than a traditional exam where the same subset of items from the pool are administered to all examinees. The limited nature of the item pool simply limits the ability to achieve equiprecision, not the ability of CAT to achieve superiority of traditional non-adaptive tests. Enhanced accuracy with a smaller number of items is still noted in the adaptive framework.

The results suggested above are based on the live testing of examinees on a paper-and-pencil examinations and on simulated candidates in the adaptive testing framework. Perhaps, further evidence for the arguments set above would be achieved

if actual examinees were tested in both frameworks, thus allowing for a comparison of within the adaptive and non-adaptive frameworks. Simulated examinee abilities are based solely upon probability theory and do not allow room for human factors which, in fact, could be present in the paper-and-pencil and CAT administrations. Further investigations along this line might implement live candidates for both frameworks.

In addition, the study could be extended by investigating results via the 2-PL and 3-PL models. Although, items did demonstrate fit to the Rasch model, comparisons of fit to the other models might have suggested that two or three parameters could have better characterized the items which constitute the item pool.

Further, each of the four paper-and-pencil exams were randomly created so to adhere to content fit. In the pilot testing of the examinations, no consideration was given to creating exams with particular difficulty distributions as was done in previous studies (Stocking, 1987). Thus further comparisons might be made between paper-and-pencil exams and CATs with limited pools, when the paper-and-pencil exams are designed to adhere to specific item difficulty distributions.

Finally, these results were studied via achievement testing; further applications of this study could examine the impact of the limited pool in certification or licensing testing which requires a definitive judgment of examinee competence needs to be established. Work in this area could be examined on examinations such as the NCLEX which currently implements an adaptive framework to determine licensure in the area of nursing.

In the end, work in the area of computerized adaptive testing needs to continue. Clearly, the negligible effect of the medium of presentation has been established (Mead & Drasgow, 1993) as has the superiority of CAT in terms of accuracy and efficiency

in hypothetical situations, as well as the ability of CAT to achieve these desirable psychometric qualities even when item pools are limited. As CAT becomes more widely implemented, practical shortcomings and issues need to be clearly documented and investigated so to assure that this shift in testing frameworks ultimately provides solutions to the traditional testing. Recent controversies such as the memorizing of items by candidates administered the Graduate Record Examination via CAT necessitate the consideration of such issues if there will ever be a large enough item pool to counteract cheating.

Still despite its limitations, CAT holds the potential for a testing process that more accurately assesses the width and breath of examinees' abilities (Cooper & Halkitis, 1995).

Appendix A

C++ Program to Generate Probabilities and Random Numbers

```
/*****************************************************************/
/*****************************************************************/
/** PVAL.C                                                      **/
/**                                                             **/
/**                                                             **/
/**                                                             **/
/*****************************************************************/
/*****************************************************************/
#include <stdio.h>
#include <stdlib.h>
#include <time.h>
#include <math.h>
#include <ctype.h>
#include <float.h>
#include <conio.h>
#include <string.h>
double b=0;
double d=0;
double p=0,dif=0;
char ans1[2]="Y";
char *ans2;
/*****************************************************************/
int main(void)
{
   FILE *fd, *fopen();
   char name[15];
   int r;  /* save random number for ouput */

   clrscr();
   randomize();

   printf ("Please enter the file name -=>");
   scanf("%s",name);
   fd = fopen (name,"a");

   do
     {
      printf("\nEnter 'B'?\n");
      scanf("%lf",&b);
      printf("\nEnter 'D'?\n");
      scanf("%lf",&d);
      dif = (b - d)*1.7;
      /*printf("%f**%f=%f\n",2.71828,dif,pow(2.71828,dif));*/
```

```
      p = pow(2.71828,dif)/(1 +pow(2.71828,dif));
      printf("\n\n\n\n");
      printf("                (b-d)*1.7\n");
      printf("           2.71828\n");
      printf("     P = ----------\n");
      printf("                 (b-d)*1.7\n");
      printf("         1 + 2.71828\n\n\n");
      printf("P = %f\n",p);
      printf("\n\n");
      r = rand() % 101;
      printf("Random Number in the 1-100 range: %d\n",r);
      fprintf (fd,"P = %f Random # = %d\n", p, r);
      printf("\nContinue? (Y or N)?\n");
      scanf("%s",ans1);
      ans2=strupr(ans1);
      if (memcmp(ans2,"Y",1)==0)
        clrscr();
     } while(memcmp(ans2,"Y",1)==0);
   close (fd);
};
```

Appendix B

Paradox Program to Simulate Ability and Standard Error of Measurement

```
;////////////////\\\\\\\\\\\\\\\\/////////////////////\\\\\\\\\\\\\\\\/////////\
;
; "θ" is ascii 233
;
;------------------------------------------------------------------------------
;
proc give_test()
 private init_beta.n,choice.a, continue.l, i_rec.r

 style
 @0,0 ?? spaces(160)
 @0,0 ?? "Enter the initial value of Beta for the false items: "
 style attribute 79
 ?? " "
 accept "N" default .3 to Init_beta.n
 if not retval then return
 endif

 empty "item"

 edit "item"
 ; add the two bogus records
  [Item]               = -1
  [Item Difficulty]     = init_beta.n
  [Correct Response?]   = "N"
  [Student Ability]     = init_beta.n
  [Student Std Error]   = 0 ; ?
  [pθ]                 = .5
  [qθ]                 = .5
  [pqθ]                = .25
  [H limit]            = 0 ; ?
  down
  [Item]               = -2
  [Item Difficulty]     = init_beta.n
  [Correct Response?]   = "Y"
  [Student Ability]     = init_beta.n
  [Student Std Error]   = 0 ; ?
  [pθ]                 = .5
  [qθ]                 = .5
  [pqθ]                = .25
  [H limit]            = 0 ; ?
  down
```

```
  pickform "F"
  [item] = recno()

  while true
  style attribute 31

  wait record
    prompt "Press F2 to process this record and Esc to quit.  F7 to View the table.",
""
    until "F2","Esc", "F7"
    choice.a = retval
    switch

      case choice.a = "Esc" : do_it! return

      case choice.a = "F7"  : view_table()

      case choice.a = "F2"  : copytoarray i_rec.r
                        continue.l = process_choice.l()
                        imagerights readonly
                        if continue.l then
                          message "Going to next item" sleep 1000
                          imagerights
                          Pgdn ;; add a new record
                          [item] = recno()
                        else
                          message "End of test"
                          return
                        endif

    endswitch

  endwhile

endproc


proc view_table()
  private rec.n

    rec.n = recno()
    imagerights readonly
    formkey
    wait table
      prompt "Viewing table,  Press Esc to return to form"
      until "Esc", "F7"
      if retval = "Esc" or retval = "F7" then
        imagerights
```

```
      moveto record rec.n
      formkey
    endif

endproc


;-------------------------------------------------------------------------


proc process_choice.l()
  private old_theta.n, new_theta.n

   if NOT isassigned( old_theta.n ) then
     old_theta.n = init_beta.n
   endif

  ;call calculation
  new_theta.n = calc_theta.n( old_theta.n )
  old_theta.n = new_theta.n
  ;; later add logic for end of test
  return true

endproc


proc calc_theta.n( old_theta.n )
  private right.n, n_items, beta.r, i, sum_p.n, sum_q.n, sum_pq.n,
        h, h_limit.n, iterations.n, se, new_theta.n, old_theta.n


  right.n = 0 ;;
  n_items = nrecords( "item" )
  array beta.r[ n_items ]
  i       = 0
  h       = 1 ; an initial value
  H_LIMIT.N = .001
  iterations.n = 0

  style attribute 30
  @14,0 clear EOS ;; set screen position

  imagerights

  ; count number of correct so far
  scan
    i = i + 1
    if [correct response?] = "Y" then
```

```
    right.n = right.n + 1
  endif
  beta.r[i] = [item difficulty]
endscan

;;;;;
;;;;; start iterations to deterimine θ with h < |.001|
;;;;;
while abs( h ) > H_LIMIT.N

  sum_p.n  = 0
  sum_q.n  = 0
  sum_pq.n = 0
  iterations.n = iterations.n + 1

  scan

    if old_theta.n = init_beta.n and [#] < 3 then
      ; don't compute p & q for the first two false items
      ; the first time around.
    else
      p.n  =  p.n( old_theta.n, [Item difficulty] )  ; beta.r[n_items] )
      [pθ] = p.n
      [qθ]  =  1 - p.n
      [pqθ] =  ( 1 - p.n ) * p.n
    endif

    sum_p.n = sum_p.n + [pθ]
    sum_q.n = sum_q.n + [qθ]
    sum_pq.n = sum_pq.n + [pqθ]

  endscan

  h = ( right.n - sum_p.n ) * 1.7 / ( -1 * sum_pq.n * 2.89 )

  se = 1 / sqrt( sum_pq.n * 2.89 )

  new_theta.n = old_theta.n - h

  ? "    ", "H value=", format( "W9.6", h ), "   N=",iterations.n, "   Old_θ=",
format( "W9.6", old_theta.n ), "   New θ=",format( "W9.6", new_theta.n )

  old_theta.n = new_theta.n

endwhile

end
```

```
  [student ability]    =  new_theta.n
  [student std error]  =  se
  [h limit]           =  h
  [n]                 =  iterations.n

  ? "    ", "Standard Error=", format( "W9.6", se )
  ? "any key to continue"

  c=getchar()

  return new_theta.n

endproc


proc p.n( theta.n, beta.n )
  private proc.a, p.n, x, ex

  proc.a = "p.n"
  ; trap for an error here!! ( test theta.n - beta.n ) > ??
  ; e^X bombs if X > 709
  x = (theta.n - beta.n)*1.7
  if x > 709 then                  ; There is no sound reason for doing this!
    message "EXPonent too large"   ; It's just that there has to be some
    x = 708                        ; sort of guard against this happening
    sleep 1000                     ; in the field!
  endif
  ex = exp( x )
  p.n = ex / ( 1 + ex )
  return p.n

endproc
;///////////\\\\\\\\\\\\\\\\\\\\\\\///////////////////\\\\\\\\\\\\\\\\\\\\

cursor off

if version() > 3.5 then
  execute "setuimode compatible"
endif

clear

give_test()
clearall
if version() > 3.5 then
  execute "setuimode standard"
endif
quit "Bye!"
```

References

Anastasi, A. (1985). Mental measurement: Some emerging trends. In J. V. Mitchell (Ed.), *The ninth mental measurements yearbook*. Lincoln, NB: The Buros Institute of Mental Measurements.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley.

Bock, R D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 37*, 29-51.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation in a microcomputer environment. *Applied Psychological Measurement, 6, 4,* 431-444.

Borland International (1985). *Paradox version 4.0*. Scotts Valley, CA: Borland International. Borland International (1993). *C + + version 4.0*. Scotts Valley, CA: Borland International.

Bryson, R. (1971). *A comparison of for methods of selecting items for computer-assisted testing* (Technical Bulletin STB 72-8). San Diego, CA: Naval Personnel and Training Research Laboratory.

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

Cliff, N, Cudeck, R., McCormick, D. J. (1979). Evaluation of implied orders as a basis for tailored testing with simulation data. *Applied Psychological Measurement, 3, 4*, 495-514.

Cooper, C. & Halkitis, P. N. (1995). This test is for you. *Wired, 3, 1*, 64-68.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart, and Winston Inc.

Cronbach, L. J. (1990). *Essentials of psychological testing*. Harper Collins: New York.

Cudeck, R. (1985). A structural comparison of conventional and adaptive of the ASVAB. *Multivariate Behavioral Research, 20,* 305-322.

De Ayala, R. J. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. *Educational and Psychological Measurement, 49,* 789-805.

De Ayala, R. J., Dodd, B. G. & Koch W.R. (1990). A simulation and comparison of flexilevel and Bayesian computerized adaptive testing. *Journal of Educational Measurement, 27, 3,* 227-240.

Divgi, D. R. (1989). Estimating reliabilities of computerized adaptive tests. *Applied Psychological Measurement, 13, 2,* 145-149.

Dodd, B. G., Koch, W. R., & De Ayala,R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13, 4,* 129-143.

Green, B. F, et al (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21, 4,* 347-360.

Halkitis, P. N. (1993). Computer-adaptive testing algorithm. *Rasch Measurement, 6, 4,* 254-255.

Halkitis, P. N. (1992). Mean square significance and sample size. *Rasch Measurement, 6, 3,* 227-228.

Halkitis, P. N. & Leahy, J. M. (1993). Computerized adaptive testing: The future is upon us. *Nursing & Health Care, 14, 7,* 378-385.

Hambleton, R. K. & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.) *New horizons in testing; Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Hambleton, R. K. & Cook, L. L. (1977). Latent trait models and their use. *Journal of Educational Measurement, 14, 2,* 75-96.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12, 3,* 38-47.

Hambleton, R. K. & Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory: Volume 2.* Newbury Park, CA: Sage Publications.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8,1,* 35-41.

Hays, W. L. (1988). *Statistics, Fourth edition.* Fort Worth, TX: Holt, Rinehart and Winston, Inc.

Henly, S. et al (1989) Adaptive and conventional versions of the DAT: The first complete test battery comparison. *Applied Psychological Measurement, 13, 4,* 363-371.

Jensema, C. J. (1972). An application of latent trait mental test theory (Doctoral dissertation, University of Washington, 1972). *Dissertation Abstracts International, 24,* 633. (University Microfilms No. 72-20,871).

Jensema, C. J. (1974). The validity of Bayesian tailored testing. *Educational and Psychological Measurement, 34*, 757-766.

Johnson, M. F. & Weiss, D. J. (1980). Parallel forms reliability and measurement accuracy comparison of adaptive and conventional testing strategies. In D. J. Weiss (Ed.) *Proceedings of the 1979 computerized adaptive testing conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Kiley, G. L., Zara, A. R., & Weiss, D. J. (1983, January). *Alternate forms reliability and concurrent validity of adaptive and conventional tests with military recruits.* Report submitted to Navy Personnel and Development Center, San Diego, CA.

Kingsbury, G. G & Houser, R. L. (1993). Assessing the utility of item response models: Computerized adaptive testing. *Educational Measurement: Issues and Practice, 12, 1*, 21-27.

Kingsbury, G. L & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2, 4*, 359-375.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Norwell, MA: Kluwer Academic Publishers.

Lewis, C. & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14, 4,* 431-444.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement, 21, 3*, 239-243.

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

Lord, F. M. (1977). A broad ranged tailored testing of verbal ability. *Applied Psychological Measurement, 1, 1, 95-100.*

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14, 2*, 117-138.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-264.

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

Lord, F. M. (1953). The relation of a test score to the trait underlying a test. *Educational and Psychological Measurement, 13*, 517-548.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph, 7.*

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Mass: Addison-Wesley.

Maurelli, V. & Weiss, D.J. (1981). *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries* (Research Rep. No. 81-4). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

McBride, J. R. (1977). Some properties of Bayesian adaptive ability testing strategy. *Applied Psychological Measurement, 1, 1,* 121-140.

McBride, J. R. (1980). Adaptive verbal ability testing in a military setting. In D. J. Weiss (Ed.) *Proceedings of the 1979 computerized adaptive testing conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

McBride, J. R, & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing; Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

McBride, J. R. & Sympson, J. B. (1985). The computerized adaptive testing system development project. In D. J. Weiss (Ed.) *Proceedings of the 1982 item response theory and computerized adaptive testing conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114, 3*, 449-458.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3, Item analysis and test scoring with binary logistic models.* Chicago: Scientific Software.

Moreno, K. E. et al (1984). Relationship between corresponding armed services vocational aptitude battery (ASVAB) and Computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8, 2,* 155-163.

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

Panchapakesan, N. (1969). *The simple logistic model and mental measurement.* Unpublished doctoral dissertation, University of Chicago.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmark Paedagogiske Institut (republished in 1980 by University of Chicago Press).

Samejima, F. (1977). A use of information function in tailored testing. *Applied Psychological Measurement, 1, 2,* 233-247.

Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph, 18.*

Sheehan, K. & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16, 1,* 65-76.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review, 36, 3/4,* 263-277.

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.) *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychomterika, 47, 2,* 175-186.

Thissen, H. & Mislevy (1990). Item response theory, item calibration and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Earlbaum Associate, Inc.

Thomas, T. J.(1990). Item-presentation controls for multidimensional item pools in computerized adaptive testing. *Behavior Research Methods, Instruments & Computers, 22, 2,* 227-252.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York: Science Press.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11,* 1-13.

Urry, V. W. (1977). Tailored testing: A successful application of item response theory. *Journal of Educational Measurement, 14, 2,* 181-196.

Urry, V. W. (1971). *Individualized testing by Bayesian estimation* (Research Bulletin 0171-177). Seattle: University of Washington, Bureau of Testing.

Vicino, F. L. & Hardwicke, S. B. (1984). *An evaluation of the utility of large scale computerized testing.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12, 1*, 15-20.

Wainer, H. (1990). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24, 3,* 185-201. Lawrence Erlbaum Associates.

Ward, W. C. (1985). Measurement research that will change test design for the future. In *The Redesign of Testing for the 21st Century, Proceedings of the 1985 Invitational Conference*. Princeton, NJ: ETS.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53, 6,* 774-789.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21, 4,* 361-375.

Weiss, D. J. & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8, 3,* 273-285.

Wright, B. D. (1977). Misunderstanding the Rasch model. *Journal of Educational Measurement, 14, 3*, 219-226.

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

Wright, B. D. & Douglas, G. A. (1975). *Best test design and self-tailored testing*. Research Memorandum No. 19, Statistical Laboratory, Department of Education, University of Chicago.