

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

UMI[®]

A

**THE IMPLICATIONS OF THE USE OF
NON-OPTIMAL ITEMS IN A
COMPUTER ADAPTIVE TESTING
(CAT) ENVIRONMENT**

by

Debra J. Grodenchik

**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.**

2002

UMI Number: 3047223

**Copyright 2002 by
Grodenschik, Debra Joan**

All rights reserved.

UMI[®]

UMI Microform 3047223

**Copyright 2002 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.**

**ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346**

Copyright

2002

DEBRA J. GRODENCHIK

All Rights Reserved

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.

5/1/02 David Rindskopf
Date Chair of Examining Committee

5/1/00 Alan J. Gross
Date Executive Officer

Alan Gross, Ph.D.

David Rindskopf, Ph.D.

Carol Tittle, Ph.D.
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

THE IMPLICATIONS OF THE USE OF NON-OPTIMAL ITEMS IN A COMPUTER ADAPTIVE TESTING ENVIRONMENT

by

Debra J. Grodenchik

Adviser: David M. Rindskopf, Ph.D.

This study describes the effects of manipulating item difficulty in a computer adaptive testing (CAT) environment. There are many potential benefits when using CATs as compared to traditional tests. These include increased security, shorter tests, and more precise measurement. According to IRT, the theory underlying CAT, as the computer continually recalculates ability, items that match that current estimate of ability are administered. Such items provide maximum information about examinees during the test. Herein, however, lies a potential problem. These optimal CAT items result in an examinee having only a 50% chance of a correct response. Some examinees may consider such items unduly challenging. Further, when test anxiety is a factor, it is possible that test scores may be negatively affected.

This research was undertaken to determine the effects of administering easier CAT items on ability estimation and test length using computer simulations. Also considered was the administration of different numbers of initial items prior to the start of the adaptive portion of the test, using three different levels of measurement precision.

Results indicate that regardless of the number of initial items administered, the level of precision employed, or the modifications made to item difficulty, the approximation of estimated ability to true ability is good in all cases. Additionally, the standard deviations of the ability estimates closely approximate the theoretical levels of precision used as stopping rules for the simulated CATs.

Since optimal CAT items are not used, each item administered provides less information about examinees than optimal CAT items. This results in longer tests. Fortunately, using easier items that provide up to a 66.4% chance of a correct response results in tests that only modestly increase in length, across levels of precision. For larger standard errors, even easier items (up to a 73.5% chance of a correct response) result in only negligible to modest increases in test length.

Examinees who find optimal CAT items difficult or examinees with test anxiety may find CATs that implement easier items enhance the already existing benefits of CAT.

ACKNOWLEDGEMENTS

This paper is dedicated in loving memory to my father, the late Prof. Ira D. Rothberg, who was taken from this world much too early. In many ways he influenced the paths I have chosen in life, especially this one.

Thanks go to my advisor, Dr. David Rindskopf and the other members of my committee, Drs. Alan Gross and Carol Tittle, for their guidance, support and encouragement throughout my time at the Graduate Center. Thank you, as well, to my outside readers Drs. Bernie Gorman and Louis Primavera.

A multitude of thanks goes to Penny D'Amato who helped me write the original program for my field research. This program served as the basis for the CAT simulations in this paper.

Thank you to all of my colleagues at NCC whose support and encouragement helped me, finally, make it through to the end of this long, arduous, often lonely, yet very rewarding process. Thank you for letting me share my frustrations as well as my accomplishments with you. Special thanks go to those of you who watched me practice and those who offered to read, especially Dr. Jane Maher who actually read portions of this dissertation and made helpful comments and recommendations.

Thank you Dr. Victor Tupitsyn, Professor Emeritus from Pace University. It was your unwavering confidence in me, so many years ago, which planted those early seeds. They grew into the confidence I needed to have in myself to embark on this voyage. I will always remember your kind words.

Many thanks to my family for their love and support, not to mention their computers! Allan, thanks for coming; it was nice to have you there. Gram, I know you can see how far I have come. I know in my heart that you will be with me in spirit at my commencement. Many thanks to my dear friends, especially Ellen, who were always ready to listen. I love you all!

Finally to the two most important people in my life... David, my big guy, thank you for understanding why mommy often needed time to do her school work. You never complained and you were always there with those special hugs and kisses I so often needed. I love you more than you could ever possibly imagine. And last, but certainly not least, Barry, I don't think I could ever have made it through without you. Thank you for never complaining about my endless hours in the library and at the computer. Thank you for your confidence in me, your undying love and support. Okay, now that I'm done, what's next?

TABLE OF CONTENTS

Copyright	ii
Approval	iii
Abstract	iv
Acknowledgements	vi
Table of Contents	viii
List of Tables	x
List of Figures	xi

CHAPTER

I. INTRODUCTION	1
II. REVIEW OF THE LITERATURE	5
A. Classical Test Theory	5
B. Item Response Theory	8
C. Item Response Theory Models	12
D. Adaptive (Tailored) Testing	23
E. Computerized Adaptive Testing	26
F. Possible Psychological and Psychometric Effects of CAT	28
III. METHOD.	35
A. Rationale for Research	35
B. Overview of Procedure.	36
C. Theoretical Examination	38
D. Participants	41

E. CAT Simulation	42
F. Analysis	45
IV. RESULTS	47
A. Manipulation of Item Difficulty	47
B. Number of Initial Test Items Administered	67
C. Ability Estimates	74
V. SUMMARY AND DISCUSSION	75
A. Ability Estimates	75
B. Number of Initial Test Items Administered	75
C. Manipulation of Item Difficulty	76
D. Implications	80
E. Future Research	82
Appendix	
A. Quickbasic program to simulate CAT test for generating ability estimates and corresponding standard errors for these estimates	84
B. Average number of test items and corresponding standard deviations	87
C. Average ability estimates and corresponding standard deviations	96
References	105

LIST OF TABLES

1. The effect of δ on $P(\theta)$	37
2. Item Information, number of test items and corresponding increase over optimal CAT condition (% increase) given modification to item difficulty	40
3. Average number of test items (n) and percent increase over optimal CAT items with s.e.(θ) of 0.2	58
4. Average number of test items (n) and percent increase over optimal CAT items with s.e.(θ) of 0.35	59
5. Average number of test items (n) and percent increase over optimal CAT items with s.e.(θ) of 0.5	60
6. Average number of test items (n), s.d.(n) and corresponding percent with s.e.(θ) of 0.2	64
7. Average number of test items (n), s.d.(n) and corresponding percent with s.e.(θ) of 0.35	65
8. Average number of test items (n), s.d.(n) and corresponding percent with s.e.(θ) of 0.5	66
9. Average number of test items (n) with s.e.($\hat{\theta}$) < 0.2	68
10. Average number of test items (n) with s.e.($\hat{\theta}$) < 0.35	68
11. Average number of test items (n) with s.e.($\hat{\theta}$) < 0.5	69

LIST OF FIGURES

1.	ICCs for the 1 – parameter logistic model where:	
	(i) $b = -1.5$ (ii) $b = 0.5$ (iii) $b = 1.5$	16
2.	ICCs for the 2 – parameter logistic model where:	
	(i) $b = -1.5$ and $a = 0.5$ (ii) $b = 0.5$ and $a = 2.0$	
	(iii) $b = 1.5$ and $a = 1.5$	17
3.	ICCs for the 3 – parameter logistic model where:	
	(i) $b = -1.5$, $a = 0.5$ and $c = 0.1$ (ii) $b = 0.5$, $a = 2.0$ and $c = 0.25$	
	(iii) $b = 1.5$, $a = 1.5$ and $c = 0.18$	18
4.	Information Function for the Rasch model with $D = 1.7$ and $b = -1.5$. .	20
5.	Information Function for the Rasch model with $D = 1.7$ and $b = 1.0$. .	21
6.	Average number of test items (n) with 3 initial test items and	
	$s.e.(\hat{\theta})$ of 0.2	48
7.	Average number of test items (n) with 5 initial test items and	
	$s.e.(\hat{\theta})$ of 0.2	49
8.	Average number of test items (n) with 7 initial test items and	
	$s.e.(\hat{\theta})$ of 0.2	50
9.	Average number of test items (n) with 3 initial test items and	
	$s.e.(\hat{\theta})$ of 0.35	51
10.	Average number of test items (n) with 5 initial test items and	
	$s.e.(\hat{\theta})$ of 0.35	52

11. Average number of test items (n) with 7 initial test items and	
s.e.($\hat{\theta}$) of 0.35	53
12. Average number of test items (n) with 3 initial test items and	
s.e.($\hat{\theta}$) of 0.5	54
13. Average number of test items (n) with 5 initial test items and	
s.e.($\hat{\theta}$) of 0.5	55
14. Average number of test items (n) with 7 initial test items and	
s.e.($\hat{\theta}$) of 0.5	56
15. Average number of test items (n) for s.e.($\hat{\theta}$) < 0.2	71
16. Average number of test items (n) for s.e.($\hat{\theta}$) < 0.35	72
17. Average number of test items (n) for s.e.($\hat{\theta}$) < 0.5	73

Chapter I

INTRODUCTION

For centuries, tests, in a variety of forms, have been employed for a variety of reasons. In more recent decades, mass test administration became possible when electronic scoring machines became available. Even more recently, computers entered the picture and computerized test administration became possible. The use of computers in testing became widespread once affordable microcomputers became widely available. This form of testing was more convenient in many ways, eliminating the need for massive numbers of paper versions. Scoring could be incorporated into the computerization of these tests. But these tests were merely electronic versions of a paper-and-pencil traditional test.

Ordinary computerized tests, along with paper-and-pencil tests, have several problems. In a traditional computerized testing situation, all examinees are administered the same exam. The examinees may finish at different times, but all examinees have seen all of the questions included on that exam. This also means that high-ability examinees will probably answer easy questions correctly, but this would provide little information about their exact ability level. Low-ability examinees will be exposed to difficult questions that they probably cannot correctly answer. Average-ability examinees see both of the extremes; they can probably correctly answer the easy questions with no difficulty, but also have very little chance of correctly answering the difficult questions. Traditional

tests, whether administered via paper-and-pencil or on the computer, basically present at least some items to all individuals that provide little or no information about that individual examinee.

Another problem with traditional testing situations is that most exams are written for average-ability examinees. There are generally very few easy/very easy and very few difficult/very difficult questions found on these exams. Therefore, this leads to a situation in which the low-ability examinees are clustered together at the low end, and high-ability examinees are clustered together at the high end of the ability continuum, without providing a means to distinguish among the individuals who place in each of these two groups. In effect, some sort of floor and ceiling effects are present. That is, traditional tests have very little ability to discriminate among individuals at each of the extremes of the ability continuum. One would expect to see differences among examinees at both the high end and the low end of the ability continuum, but traditional test scores generally do not reflect these differences. This is because very few test items are geared toward high-ability, as well as low-ability, examinees.

With the development of item response theory (IRT) came the ability to separate test characteristics (such as specific test items administered and test length) from examinee characteristics (such as ability level). IRT made it possible to administer different tests to individuals of similar ability, and the resulting test scores would then reflect these abilities even though the examinees were administered different exams. Additionally, test items could be calibrated with one sample of examinees and administered to a different group of

examinees and no information would be lost. In other words, tests designed and calibrated according to IRT specifications are independent of the examinees that were the basis of that calibration. Additionally the scores achieved on such a test are independent of the specific test items administered to any individual examinee, both in number and difficulty.

The marriage of computerized testing and item response theory gave rise to computerized adaptive testing (CAT). In a CAT an individual sits at a computer station to take a computerized test, but this test differs greatly from any ordinary computerized test. The unique feature about computer adaptive tests is that they are tailored to the individual examinee's ability level. As an examinee answers test items, the computer is programmed to appropriately choose the next item, one that will provide the most information about that examinee, based on the examinee's current ability estimate. With each successive response, the computer is programmed to recalculate the examinee's ability level given the difficulty of the items answered and the correctness of those answers. For example, if the tenth item is answered correctly, the ability estimate will increase from the previous estimate based on the first nine test items, and the eleventh item will be more difficult than the tenth item. If the tenth item is answered incorrectly, the ability estimate will decrease and the eleventh item will be easier.

In IRT, the ideal items to administer to any examinee are those that provide maximum information about that examinee. According to IRT, the items that provide this maximum level of information have levels of difficulty that exactly match an individual's true ability level. Herein lies a potential problem. Based on

the IRT model to be employed in the following study (the Rasch model), when item difficulty is equal to an individual's ability level, that individual will only have a 50% chance of correctly answering that item. (The theory behind this will be addressed later.) On any given CAT, examinees will only correctly answer approximately 50% of the test items administered. This may be a situation that low-ability examinees are accustomed to, although they are probably not happy about it, but it is a situation with which average and high-ability examinees are certainly not familiar. In fact, some may say that if a test item is chosen such that it only provides an individual with a 50% chance of a correct response, that question may be considered very challenging. Now imagine an entire exam composed of such items. Such an exam may cause undue stress and anxiety due to the level of difficulty of the items included on that exam.

If there are psychological consequences to administering optimal CAT items, as dictated by IRT, what are the psychometric consequences of alleviating this purported stress by administering easier items? Since scores on IRT-based tests are independent of the items administered, changing the difficulty level of test items should not have any noticeable impact on the final ability estimate. By administering easier items, it should take more test items to arrive at that final ability estimate, as compared to administering optimal items. How much easier can items become before a sizeable impact on test length is noticed? What will happen as the level of precision desired for the final ability estimate changes? This study will explore these questions.

Chapter II

LITERATURE REVIEW

A. Classical Test Theory

Classical test theory has been in place for decades in its application to the field of psychological testing. In fact, Weiss & Kingsbury (1984) assert that, until the 1970's, all educational tests, other than certain individualized tests, were administered according to classical measurement methods. Large scale mental testing began in 1917 with the military administering psychological exams to recruits using a carefully developed exam called Army Alpha. Over the years this test was modified and is currently administered as the Armed Services Vocational Aptitude Battery (ASVAB) (Wainer, 1990). In the 1920's the College Board introduced the Scholastic Aptitude Test (SAT), used to assist with college admissions and scholarship decisions, that demonstrated a resemblance to the ASVAB. By the 1940's, after revisions, the SAT appeared very similar to the current form of the SAT now administered by the Educational Testing Service (ETS).

Classical test theory is based on weak assumptions, and therefore is widely applicable (Hambleton & Swaminathan, 1985; Lord, 1980; Lord & Novick, 1968). Unfortunately there are many limitations of classical measurement models. For example, item characteristics such as item difficulty and item discrimination are dependent on the particular group of examinees from which the item

characteristics are obtained. In fact, according to Hambleton and Swaminathan (1985), item statistics can be strongly influenced by both the average ability level and the range of ability levels in the sample of examinees employed to obtain the item characteristics. For example, when this sample of examinees has a higher ability level than that of the general population of examinees, the item difficulties tend to be higher than if the sample were more representative of the population. Additionally, when the sample is heterogeneous, some item statistics, such as item discrimination, tend to be higher than when the ability levels in the sample are less varied (Hambleton & Swaminathan, 1985). This is because item discrimination in the classical model is the correlation between item score and total test score: with more homogeneous samples there is a restriction of range which results in lower correlations and therefore lower item discrimination.

In classical test theory it is very difficult to compare different groups of examinees that take different tests. For example, unless tests are strictly parallel, it would be difficult to make comparisons between different groups taking different tests. The unfortunate reality is that one can never be certain that these forms are strictly parallel. Additionally, from a classical standpoint, it would be very difficult to compare different groups of examinees taking different forms of an exam where the test items represented dramatically different levels of difficulty (Hambleton & Swaminathan, 1985). For example, it would be extremely difficult to compare a student who performed well on an easy test to a student who did not perform as well on a more difficult test.

When using exams constructed with classical measurement methods, one must consider the fact that item characteristics are group dependent and examinee scores are test dependent (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991). Hambleton and Swaminathan (1985) and Lord (1953) also note that ability estimates that are obtained using classical measurement methods are not as precise for examinees who have either high or low-ability levels as compared to those examinees whose ability is closer to the average. In the classical model, maximizing variance increases reliability. It can be shown that variance ($\sigma_x^2 = pq$, $q = 1 - p$, where p is the proportion correct) is maximized when $p = 0.5$, that is, when items are of average difficulty (Crocker & Algina, 1986; Guilford, 1954). Therefore, tests developed using classical measurement methods are most appropriate for examinees of average-ability (Green, 1983).

Another limitation of classical test theory lies in the realm of reliability. According to classical test theory, reliability is defined in terms of parallel test scores. One of the assumptions of classical test theory is that we can build strictly parallel tests (Crocker & Algina, 1986). Unfortunately, as previously stated, this is difficult to achieve and one can never be sure that tests are parallel in real world applications (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985; Hambleton et. al., 1991; Lord, 1980)

Another assumption of classical test theory is that the standard error of measurement (SEM) is the same for all examinees, but Hambleton, Swaminathan and Rogers (1991) indicate that SEM is a function of test score

reliability and variance, which vary with ability level as well as homogeneity of ability. Hambleton and Swaminathan (1985) indicate that it is, in fact, common to see consistency in test scores varying with different levels of ability. Crocker and Algina (1986) agree that SEM probably varies according to true score and indicate that standard errors at the extremes are probably smaller than those associated with scores in the middle ability range.

Lastly, classical measurement models provide no information about the test-taker at the item level. That is, classical measurement models provide no method of predicting or explaining how any particular individual examinee may perform on any particular test item (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985; Hambleton et. al., 1991). This can be troublesome when trying to adapt test items to an individual examinee's demonstrated ability level. Suppose one were trying to match the difficulty of test items to an individual's estimated ability, such as in an adaptive test. The individual's probability of correctly answering any particular test item would be important information to have, but it is not available in a classical testing environment.

B. Item Response Theory

Wright and Stone (1979) indicate that in order to achieve truly objective measurement, the calibration of item characteristics must be independent of the sample of examinees used to calibrate these item characteristics, i.e., sample-free item calibration. Additionally, the measurement of ability of any given

individual must be independent of the specific test items administered during the test, i.e., test-free person measurement. IRT, which provides an alternative to classical test theory, includes these features. This property of parameter invariance is seen by many as one of the most important characteristics of IRT (Lord, 1980; Zickar, Overton, Taylor, & Harms, 1999). Lord (1980) states that the invariance of item parameters across groups of examinees is not only one of the most important characteristics of IRT, but this invariance of item and ability parameters is noted as the cornerstone of IRT as well as a major separation from classical measurement models.

IRT is an item-centered model rather than a test-centered model and is not based on parallel forms of a test (Wainer 1990). IRT provide a means of estimating ability, and the precision of this estimate, independently of the specific items administered (Hambleton et. al., 1991). Essentially, IRT helps one developing a test accomplish the quality of objective measurement as described by Wright and Stone (1979). Hambleton and Swaminathan (1985) suggest the application of IRT to help address the disadvantages of classical test theory. In addition, they suggest that IRT will help overcome other testing problems such as test equating, test design and identifying biased items.

IRT is a latent trait model (as is the classical model), and as such, describes a relationship between an observed test score and the underlying latent trait, such as ability. The underlying trait is not directly measurable but, once estimated, enables one to predict test performance (Hambleton & Swaminathan, 1985). The observable responses to test items allow one to make

inferences about any particular examinee's ability. These inferences are made through the application of a mathematical function that describes the relationship between the examinee's observed responses and the underlying latent trait (Weiss, 1983). IRT has been described as a way to mathematically explain what happens when an individual is presented with an item. Wainer and Mislevy (1990) note a strength of IRT as the ability to deal on the item level rather than be restricted to the test level as with the classical model. Pioneering work in IRT was formalized in the early 1950's by Frederick Lord, who is seen by many as the father of IRT (Hambleton & Swaminathan, 1985). Contributions to IRT, as applied to dichotomously scored items, were made by Georg Rasch and Allan Birnbaum, with the latter contributing much of the theory (Hambleton & Swaminathan, 1985; Lord & Novick, 1968).

At the heart of IRT lies the item response function, also known as the item characteristic curve (ICC). This mathematical function describes the probability of an individual correctly answering a particular item given the item characteristics and the individual's ability level, θ . One formula which determines an examinee's chances of correctly answering a particular item, $P(\theta)$, given the examinee's ability level (θ) and level of difficulty for that particular item (b_i), is $P(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}$ (D is a scaling constant). The amount of information an item provides about that examinee's ability is inversely proportional to the variance of the estimated value of ability. So, as the error term decreases, item information increases. Item information is maximized when an individual's true ability level exactly matches the level of difficulty (b) for that

item, that is, when $\theta = b$, this equation simplifies to $P(\theta) = 0.5$. According to elementary calculus this also occurs at the inflection point on the ICC, where the slope of the ICC is steepest, because item information is a function of the first derivative of $P(\theta)$. A more detailed discussion of item information will be presented in the next section.

Item response functions are specified to be monotonic (Trabin & Weiss, 1983). This means when $\theta_1 < \theta_2$, $P(\theta_1) < P(\theta_2)$, where $P(\theta)$ relates values on the latent trait, θ , (person parameter or ability level) to the probability of passing the item (Hambleton & Swaminathan, 1985; Lord, 1953, 1980; Lord & Novick, 1968). That is, as ability increases, the probability of passing an item must increase under the condition of monotonicity.

Unlike classical test theory, item response models are based on strong assumptions and, therefore, are falsifiable (Hambleton, et. al., 1991; Trabin & Weiss, 1983). There are two fundamental assumptions underlying IRT models: unidimensionality and local independence (Hambleton & Swaminathan, 1985; Hambleton et. al., 1991; Lord, 1980; Lord & Novick, 1968). The assumption of unidimensionality states that a single latent trait explains all of the variation between items. It should be noted that saying a single factor must fit the data is extremely demanding. This is almost never true with real data. Hambleton, Swaminathan & Rogers (1991) note that this assumption is considered close enough to being met when there is a single dominant factor that influences test performance.

The assumption of local independence follows directly from unidimensionality (Hambleton & Swaminathan 1985; Lord 1980). Theoretically, local independence states that item scores will be independent in groups who are the same on the latent traits (person parameters), i.e., when conditioning on person parameters (Birnbaum, 1968). That is, when controlling for person parameters, there is nothing in any item that influences the probability of passing or failing any other item. This independence of item scores, indicating no relationship among item scores, only holds for people who are identical on the person parameters. This assumption of local independence will be violated when information on one item has the potential to help with another item(s). Local independence is represented by the following mathematical identity:

$$P(p \text{ item scores} | r \text{ person parameters}) = \prod_{i=1}^p P(\text{item score } i | r \text{ person parameters}).$$

A major advantage of the assumptions of IRT over classical test theory is that they are falsifiable, so the fit of item response models may be tested. This is not possible with classical measurement models (Lord, 1980; Trabin & Weiss, 1983).

C. Item Response Theory Models

Three IRT models are commonly discussed: 1 - parameter, 2 - parameter, and 3 - parameter logistic models. (Crocker & Algina, 1986; Lord & Novick, 1968; Weiss, 1983). For the 1 - parameter logistic model, also known as the Rasch model, one parameter per item contains all characteristics of the item (Rasch, 1960; Hambleton & Swaminathan, 1985; Hambleton et. al., 1991). That

is, the one item feature that will determine performance is item difficulty, b_i . In this model $P(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}$, where $0 < P(\theta) < 1$. Here $\theta - b_i$ is the difference between the latent trait value for the person and the difficulty of the item. The item difficulty parameter, b_i , describes the point of inflection on the ICC and is usually scaled between -2.5 and 2.5 (Ree & Jensen, 1983). Recall that the point of inflection on $P(\theta)$ occurs when $b = \theta$ and this is the point where the information an item provides about θ is maximized.

D is a scaling constant, which is typically fixed at 1 or 1.7. When D is set at 1.7 there is less than a 0.01 difference between the logistic cdf and a normal cdf with a mean of zero and standard deviation of 1.7, for all values of θ (Birnbaum, 1968; Hambleton & Swaminathan, 1985). Birnbaum (1986) states because we are able, we should choose to use a scaling factor of $D = 1.7$ "to maximize agreement between quantitative details in the normal and logistic models" (p.400).

The 2 - parameter logistic model considers variations in item discrimination in addition to item difficulty. Here passing an item is related to item difficulty and the rate at which $P(\theta)$ changes. Where the slope of $P(\theta)$ on the item response curve is very steep, the item discriminates better than where the slope of $P(\theta)$ is flatter. That is, items discriminate better for people in a certain θ range where the probability of passing increases more sharply than anywhere else on the item response curve. In this model, $P(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$ (Birnbaum, 1968; Hambleton & Swaminathan, 1985; Hambleton et. al., 1991), where a_i , the item

discrimination parameter, should be positive. Typically a_i ranges from 0.5 to 2 (Ree & Jensen, 1983).

For both the 1 - parameter and 2 - parameter logistic models the item difficulty parameter, b_i , is determined by the point on the θ scale where $P(\theta) = 0.5$, which occurs when $b_i = \theta$. As b_i moves up the θ scale, that is, as items increase in difficulty, $P(\theta) < 0.5$ for larger values of θ . In the 2 - parameter logistic model, as a_i , the discrimination parameter, increases, the probability of passing an item will increase more rapidly as θ changes. For items with a high value of a_i , the probability of passing those items will go from low to high quickly, on a narrow range of θ . For items with a lower value of a_i , the probability of passing those items will go from low to high more gradually, over a wider range of θ (Hambleton & Swaminathan, 1985).

Birnbaum's (1968) 3 - parameter logistic model adds the pseudo-guessing parameter, c_i , to the prior models. In this model $P(\theta) = \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}} + c_i$, where c_i is typically between 0 and 0.3 (Birnbbaum, 1968; Ree & Jensen, 1983). This value may be zero for some items and c_i may be different for all items. In the prior two models, $P(\theta)$ approaches zero as θ approaches negative infinity, but in this model $P(\theta)$ approaches c_i as θ approaches negative infinity. That is, an individual who is very low on the θ scale may have a higher probability of passing the item than would be expected based on their θ value. This is due to

the fact that the nature of some items allow for correct answers as a result of guessing, such as with multiple choice items. In both the 1 - parameter and 2 - parameter logistic models, when $b = \theta$, $P(\theta) = 0.5$. In the 3 - parameter logistic model $P(\theta) = \frac{c+1}{2}$ when $b = \theta$ (Hambleton & Swaminathan, 1985). That is, for the 3 - parameter logistic model, $P(\theta) = 0.5$ is no longer optimal. If $c_i = 0$ for all items, a_i is the same for all items, and $D = 1$, this 3 - parameter logistic model would revert back to the most restrictive model, the Rasch model (Birnbbaum, 1968).

Figure 1 illustrates $P(\theta)$ for the 1 - parameter logistic model (the Rasch model with $D = -1.7$) when: (i) $b = -1$, (ii) $b = 0.5$, and (iii) $b = 1.5$. Figure 2 illustrates $P(\theta)$ for the 2 - parameter logistic model when: (i) $b = -1.5$ and $a = 0.5$, (ii) $b = 0.5$ and $a = 2.0$, and (iii) $b = 1.5$ and $a = 1.5$. Figure 3 illustrates $P(\theta)$ for the 3 - parameter logistic model when: (i) $b = -1.5$, $a = 0.5$ and $c = 0.1$, (ii) $b = 0.5$, $a = 2.0$ and $c = 0.18$ and (iii) $b = 1.5$, $a = 1.5$ and $c = 0.25$.

Certain models are appropriate for each type of test items. As noted above, the 3 - parameter logistic model is appropriate for multiple choice items, but it would not seem appropriate to use either the 1 - parameter or the 2 - parameter logistic model for these items because neither model contains the pseudo-guessing parameter. For these latter two models, fill-in questions or open-ended questions that are pass/fail would seem to be ideal. For the Rasch model, as previously stated, all items must have the same item discrimination

**Figure 1: ICC for 1-parameter logistic model where:
 (i) $b = -1.5$ (ii) $b = 0.5$ (iii) $b = 1.5$**

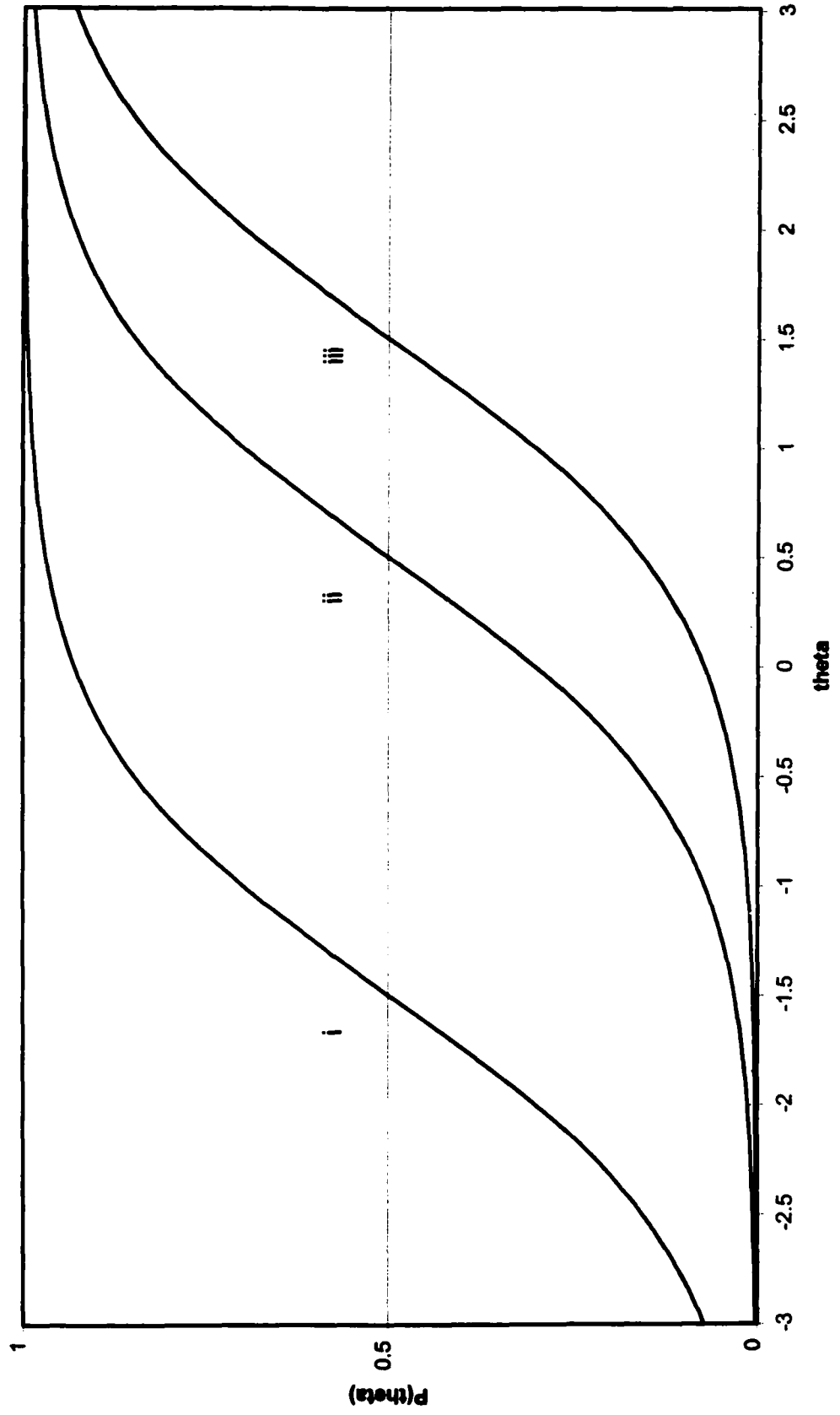


Figure 2: ICC for 2-parameter logistic model where:

(i) $b = -1.5$, $a = 0.5$ (ii) $b = 0.5$, $a = 2.0$ (iii) $b = 1.5$, $a = 1.5$

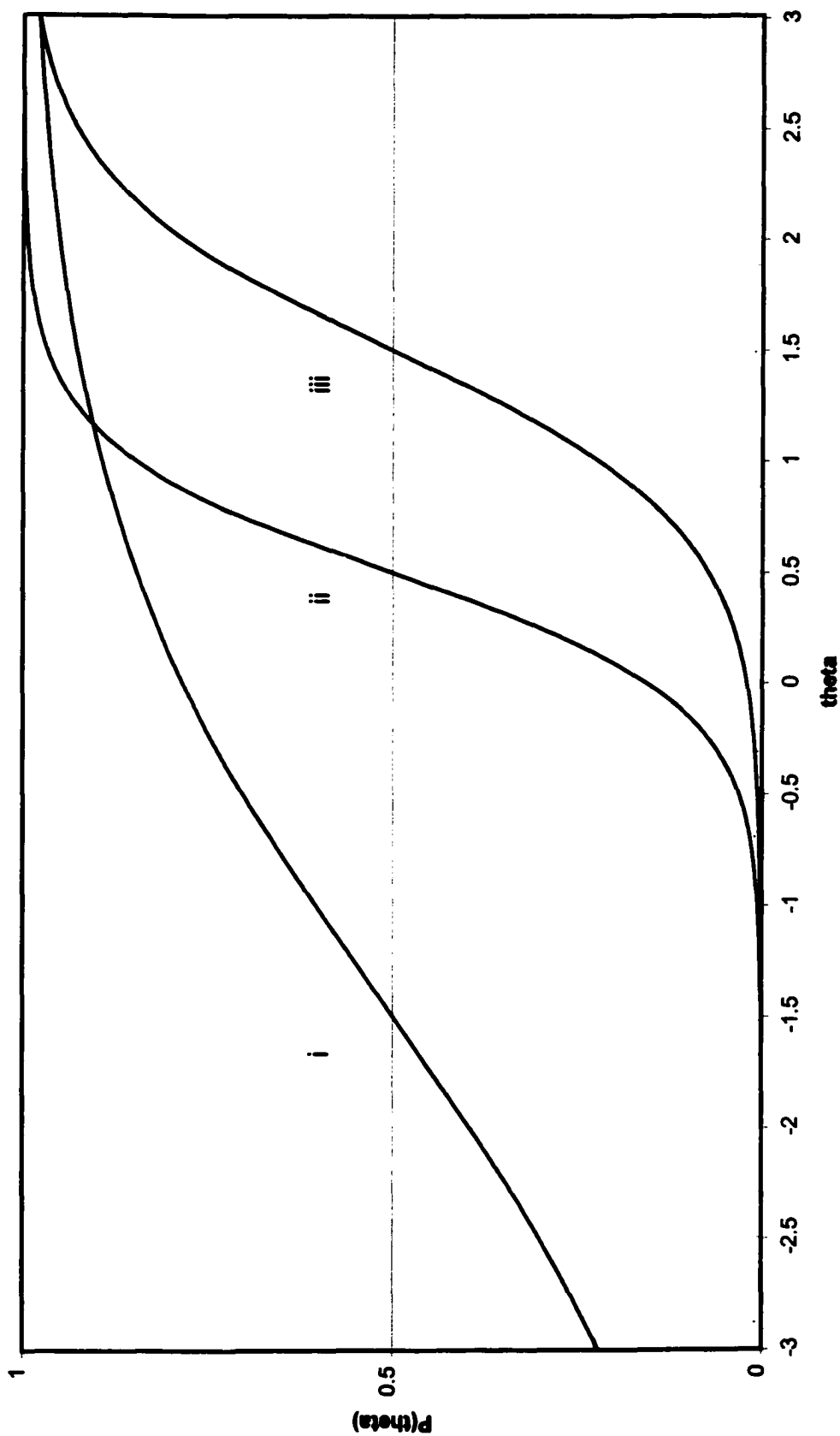
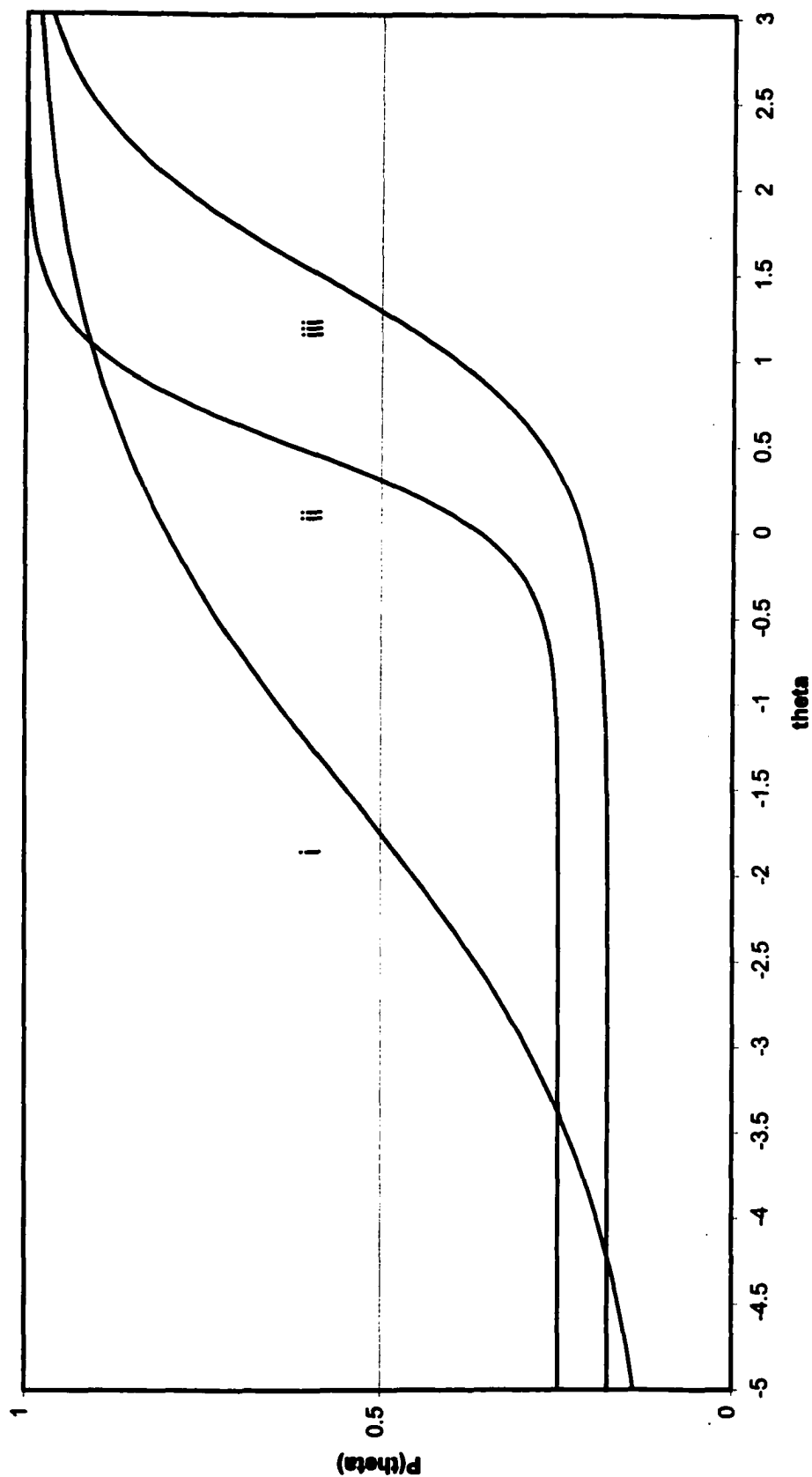


Figure 3: ICC for 3-parameter logistic model where:

- (i) $b = -1.5, a = 0.5, c = 0.1$
- (ii) $b = 0.5, a = 2.0, c = 0.25$
- (iii) $b = 1.5, a = 1.2, c = 0.18$



level. Here items could be pre-selected to have the same a , (Hambleton et. al., 1991).

IRT replaces the classical notion of "reliability" with test or item

"information" functions. The item information function is $I(\theta) = \frac{[P'(\theta)]^2}{P(\theta)Q(\theta)}$, where

$Q(\theta) = 1 - P(\theta)$ (Lord, 1980). Information refers to the amount of information an item score provides about θ . Green (1983) suggests that because information is a function of ability, adaptive testing provides a way of directly controlling test information. A given item will provide much information in some ranges of θ and less information in other ranges of θ . The item information curve peaks when $b = \theta$ (Weiss, 1985), which results in $P(\theta) = 0.5$ for the 1 – parameter and 2 – parameter logistic models. It is interesting to note that optimal items to use in the classical model also have $p = 0.5$, as previously discussed (Guilford, 1954).

Information changes depending on where in the θ range one is and is related to item discrimination and the derivative of $P(\theta)$. The steeper the ICC, the more information one has about θ (for values near b), whereas less information is available on a flatter ICC (Weiss, 1985). That is, on the steepest portion of the ICC, which occurs when θ is close to b , the slope of the ICC is larger than elsewhere on the ICC. $I(\theta)$ is a function of $P'(\theta)$, (the slope of $P(\theta)$) and, therefore, when the slope of $P(\theta)$ is large, $I(\theta)$ is large. Figure 4 provides an example of $I(\theta)$ when $b = -1.5$ using the Rasch model with $D = 1.7$. Figure 5 provides an example of $I(\theta)$ when $b = 1$ using the Rasch model with $D = 1.7$.

Figure 4: Information Function for the Rasch Model
with $D = 1.7$ and $b = -1.5$

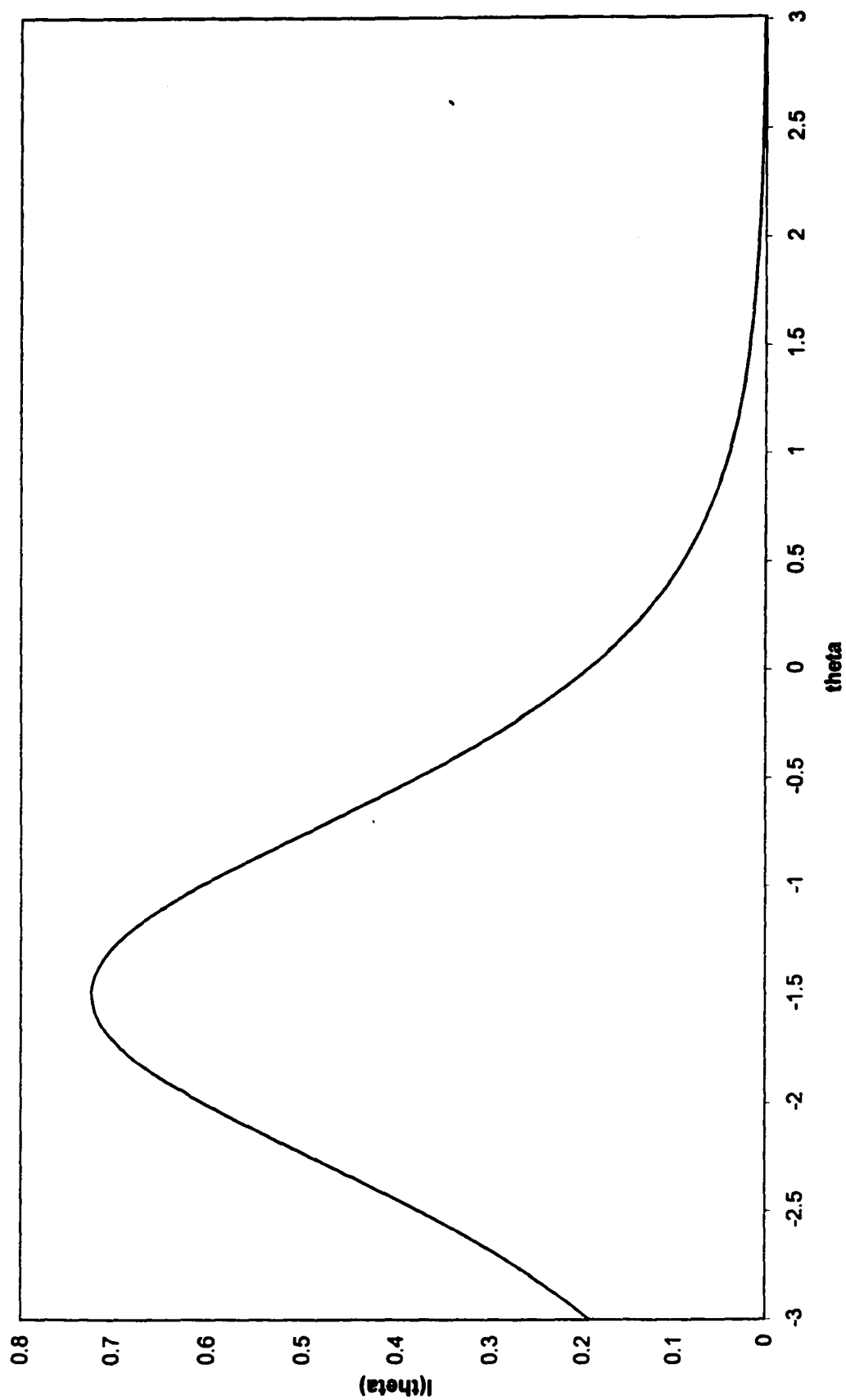
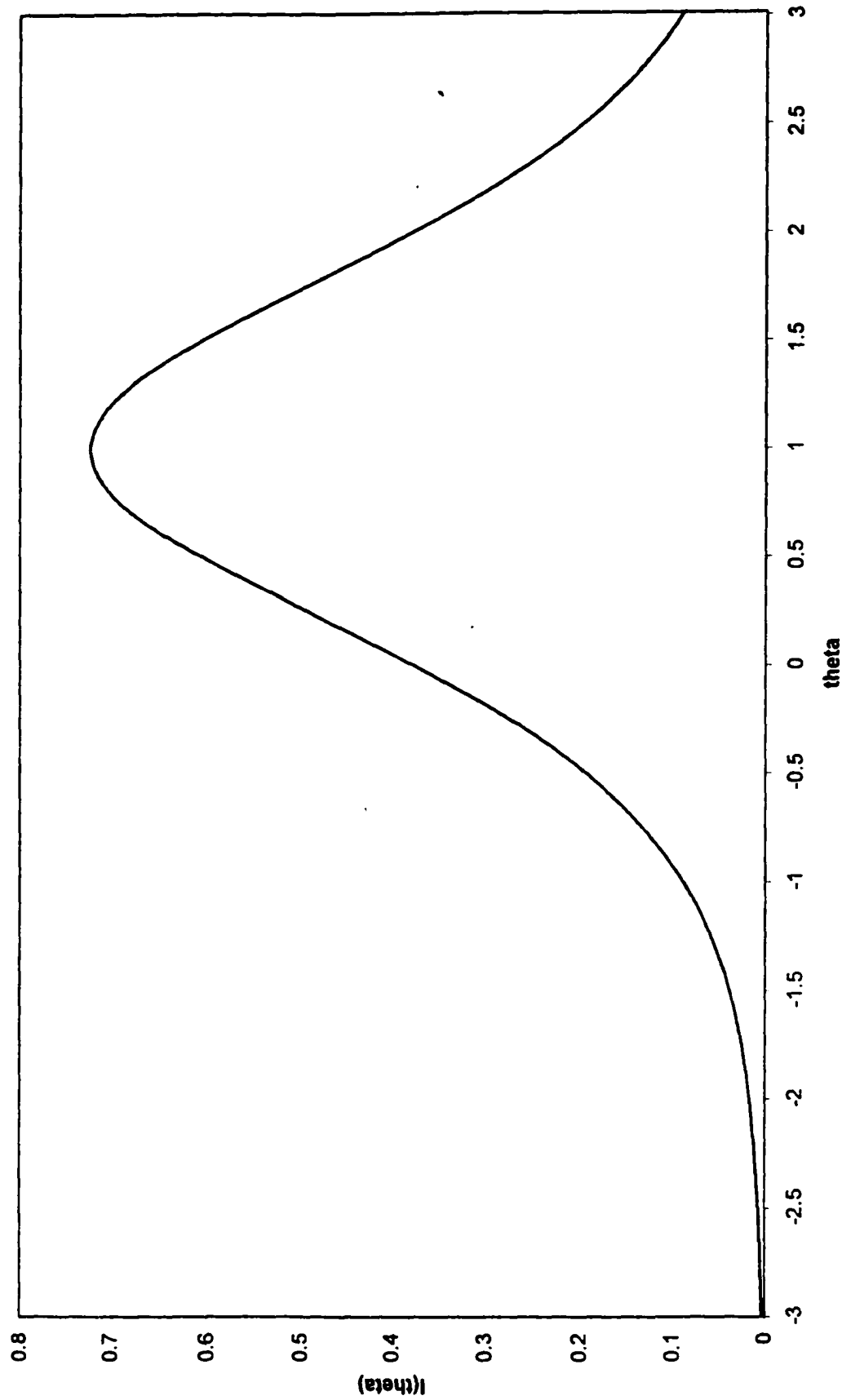


**Figure 5: Information Function for the Rasch Model
with $D = 1.7$ and $b = 1.0$**



In the Rasch model $I(\theta)$ is a function of item difficulty only. In the 2 - parameter logistic model, $I(\theta)$ is a function of item difficulty and item discrimination. As item discrimination increases, there is better information, but this is restricted to the θ range where the item discriminates well (near b). In the 3-parameter logistic model, $I(\theta)$ is a function of item difficulty, item discrimination, and guessing. With this model, there is not a great deal of discrimination at the low end of the item response curve because of guessing and, therefore, there is little information at this end of the ability continuum (Hambleton & Swaminathan, 1985).

Test information is the result of summing item information over all items included in any particular test (Lord, 1980). If items are chosen to provide maximum information for an individual examinee, test information will be maximized as a result. Because the standard error of the estimate of ability is inversely proportional to $I(\theta)$, that is, $\text{s.e.}(\theta) = \frac{1}{\sqrt{I(\theta)}}$, as information increases, error decreases (Hambleton et. al. 1991).

There is a demonstrated relationship between IRT item information and the classical standard of error measurement (Thissen, 1990). The inverse of the IRT information function yields the variances of the θ estimates for each value of θ . If one were to average over the information function with respect to θ , one would obtain the classical squared standard error of measurement, σ_e . Error is related to the classical notion of reliability, ρ , by the equation: $\sigma_e = \sigma_x \sqrt{1 - \rho}$. This simplifies to $\sigma_e = \sqrt{1 - \rho}$ when assuming a normal distribution with mean of

zero and standard deviation of one. Thissen (1990) indicated that although it is slightly unusual to set $\sigma_x = 1$ it does simplify matters. If one is considering a reliability that is equal to 0.9 then $\sigma_e = \sqrt{0.1} = 0.3162$, approximately. This is the standard error of the estimate in both traditional test theory and IRT (Thissen, 1990; Urry, 1977). In this case, when $\sigma_e = 0.3162$, $I(\theta) = 10$. When reliability drops to 0.7, measurement error increases to 0.54 and $I(\theta)$ decreases to approximately 3.33.

Hambleton, Swaminathan & Rogers (1991) offer the suggestion that while IRT is not a magic wand that can solve all the problems encountered in testing, it can become a powerful tool when appropriately applied to test construction and reporting and interpreting test results. IRT is used today by testing companies to assist with test design, assembly, scaling and calibration. IRT is also used to construct item pools for tests and to assess item bias (Hambleton et al, 1991). In addition to the application of IRT methods as mentioned above, another natural application is to adaptive, or tailored, testing. In fact, Urry (1977) describes tailored testing as "a remarkably effective application of latent trait theory" (p.181), also known as IRT.

D. Adaptive (Tailored) Testing

Hambleton and Swaminathan (1985) indicate that a particular benefit of adaptive testing procedures is that they provide more information for examinees at the high and low ends of the ability distribution than any traditional test. Lord

(1977) indicates that when a traditional test is administered to a heterogeneous group of examinees, it is impossible to determine accurate measurements for examinees at both extreme ability levels simultaneously. Traditional tests would provide maximum reliability if, for all test items on a particular test, individual examinees had a 0.5 probability of a correct response, but this never occurs in reality. This feature is possible to achieve, though, when using adaptive tests because ability estimates are independent of the subset of items administered (Hambleton & Swaminathan, 1985; Lord, 1980). Therefore, it is possible to choose items for which any given examinee has a probability of 0.5 of correctly answering these items. This is regardless of the levels of difficulty associated with these items due to the property of parameter invariance (Hambleton & Swaminathan, 1985). Recall that the property of parameter invariance states that item characteristics are independent of the ability distribution of examinees used in determining those item characteristics and examinee characteristics are independent of the test items administered to any particular examinee (Hambleton, et. al., 1991).

The earliest example of adaptive testing is Binet's intelligence testing which began in the early 1900's (Hambleton et al, 1991; Weiss, 1983; Weiss, 1985). Binet's test is said to have all the features of currently used adaptive tests, without the computer. Starting items were determined by the test administrator based on an initial estimate of the examinee's ability (Weiss, 1983). Binet's test items were scored immediately and were then used to select the next item to be administered based on the response to the previous item. The test ended when

the examinee could no longer respond to items correctly. Because of this termination procedure tests for different individuals were likely to differ in length (Weiss, 1983).

Because Binet's intelligence test was administered on an individual basis to one examinee by one examiner, it was not difficult to apply an adaptive format. For decades large-scale application of adaptive testing methods was much more difficult (Green, 1983). Costs and manpower requirements would be prohibitive if all adaptive test needed to be administered on an individual basis. In the early 1930's, professors collaborated to develop a mechanical test-scoring machine. Electronic scoring made a large impact on the testing world by allowing for large scale testing, with mass scoring, at a greatly reduced cost. Years later, with the pairing of computerized tests and IRT methods, test designers were able to create an adaptive environment, known as computerized adaptive testing (CAT), which further improved test administration (Wainer, 1990).

Urry (1977) describes the basic tailored test situation as one in which an examinee sits at a computer and is presented with test items. As items are answered correctly, item difficulty increases. As items are answered incorrectly, item difficulty decreases. With each response, the computer recalculates the examinee's ability estimate, and simultaneously, the precision of this ability estimate generally increases. The tests generally terminate when a specified level of precision has been reached. Tests may also terminate after a specified number of test items have been administered.

E. Computerized Adaptive Testing

Advances made in micro-computing with respect to hardware, software, and availability have made wide scale adaptive testing feasible (Bergstrom & Lunz, 1999; Hambleton et. al., 1991; Weiss & Kingsbury, 1984). The first high-stakes, large-scale adaptive test to be administered by computer was the CAT-ASVAB (Segall & Moreno, 1999). The CAT-ASVAB was introduced at a small number of testing locations in 1990 and is currently being administered to over 65% of all military applicants (Segall & Moreno, 1999). Development of the GRE-CAT began in 1988. The exam was introduced in 1992 as a computerized version of the paper-and-pencil test. The test became fully operational as the GRE-CAT in 1993 (Guernsey, 2000; Mills, 1999).

Other exams that are now being offered in a CAT format are the Test of English as a Foreign Language, admission tests for nursing and dental schools, and the Graduate Management Admission Test, which has been administered exclusively on the computer since 1997 (Guernsey, 2000). Additionally, Halkitis and Leahy (1993) discuss the shift from a traditional exam to a CAT format for the national nursing board exam, the National Council Licensure Examination (NCLEX). There are also plans to offer computerized versions of the Law School Admission Test and the Medical College Admission Test within the next ten years or so (Guernsey, 2000).

There are many advantages to using CAT methods. The first is shorter test length (Plake, 1995; Wainer, 1990). In fact, McBride (1985) and McBride and

Martin (1983) indicate that if adaptive tests are well constructed, resulting scores can be as reliable as conventional tests that are twice as long. Weiss (1985) found that computer adaptive tests could be as much as 50% shorter than conventional tests, in testing time, and still resulted in measurements that were more precise. This benefit is a direct result of the adaptive environment. When any individual is taking an adaptive test, questions are geared directly to their estimated ability at that particular time during the test. Because of this, higher ability students are only given the more difficult questions, weaker students are given the easier questions, while average students generally do not see questions from either extreme on the ability continuum, thus shortening the length of the test for all students.

Another advantage of CAT is improved security (Green, 1983; Olson-Buchanan & Drasgow, 1999; Wainer, 1990; Zickar, et al 1999). It is unlikely that any two students taking an adaptive test in the same area will be administered the same exact test, though this will depend on the size of the item pool and the rules employed for choosing test items. In addition, it is unlikely that any one individual taking a particular test more than once will be administered the exact same test due to individual variations from day to day.

Another advantage of computerized adaptive testing is immediate scoring which is a result of the use of computers and IRT models (McBride, 1985; Green, 1983; Wainer, 1990). Additionally, in an adaptive environment, new items can be easily tested and flawed items can be easily eliminated without a direct impact on the test itself (Wainer, 1990). Olson-Buchanan and Drasgow

(1999) note other benefits of CAT such as increased measurement precision as well as decreased labor costs. McBride and Martin (1983) and Green (1983) indicate that CAT results in fewer scoring errors because items are scored by the computer, thus avoiding any problems encountered when using separate answer sheets.

Finally, Plake (1995) suggests that computer adaptive testing allows for a more a flexible test environment. For example, in large-scale traditional testing situations large numbers of examinees gather at the same time and proceed through the testing experience from instructions through the individual sections of the test at approximately the same time. In a computer adaptive testing environment exams could be offered more often to much smaller groups of examinees with flexible start times. Instructions would be available on the computer and examinees could proceed through the different sections of the test at their own pace.

F. Possible Psychological and Psychometric Effects of CAT

As previously stated, measurements for examinees are most appropriate and effective when the test items administered are neither too difficult nor too easy. Psychometrically this translates to examinees only knowing approximately half of the answers to the questions on their individualized tests (Lord, 1980). McBride and Martin (1983) state that CAT eliminates the problem of administering tests that are excessively difficult. But one would only expect low-

ability examinees to see traditional tests as excessively difficult. High-ability examinees may actually see most traditional exams as relatively easy if they are accustomed to correctly answering 90% or more of the questions on most traditional tests.

Kingsbury and Houser (1999) explain that CAT allows examinees to take tests that are challenging, but not frustrating. Wainer (1990) presents hypothetical situations in which students comment that their adaptive tests were neither too difficult nor too easy, but one may wonder if administering items in which an individual only has a 50% chance of responding correctly may be unduly challenging. One has to wonder what psychological effect results from an examinee taking a challenging exam. And one must consider that if there are such psychological effects, do these in turn have a psychometric effect? Andrich (1995) suggests that perhaps items that provide maximal statistical information may result in too low a success rate, i.e. 50% correct. It is suggested that such a low correct response rate may make it difficult to maintain examinee interest, motivation and enthusiasm throughout the test. Because of this, Andrich (1995) suggests the effects of different success rates be researched.

Lord (1980) reminds us that a measuring instrument should not alter the trait that it is meant to measure. Is it possible that computer adaptive tests may be doing just this? Bergstrom, Lunz, & Gershon (1992) assert that a test in which individuals only have a 50% chance of responding correctly to the items is often considered hard by conventional test standards. Lunz & Bergstrom (1994) say that this particular testing situation may, in fact, not be acceptable to a

population that has substantial experience with traditional paper-and-pencil, multiple-choice tests. Guernsey (2000) suggests that examinees who are accustomed to classical scoring schemes will have to work hard to get used to the new CAT scoring system. This is because classical methods are based on the number correct, therefore 50% correct will generally be thought of as a failing grade,

Jacobson (1993) discussed the fact that top students, who are typically used to answering most standardized test questions correctly, could find CAT items more challenging, "if not also a bit unnerving" (p. A23). Ward's (1988) research on a College Board computerized placement test included reporting students' reactions to computerized tests. The results indicate that 48% of the students in the study believed the computerized tests were more challenging than traditional paper-and-pencil tests, while only 13% found the computerized tests to be less challenging.

Another issue is that being able to answer only 50% of the items correctly may be less crucial to weaker students who are accustomed to doing poorly on tests. This may be more essential to brighter students who are accustomed to answering a vast majority of test questions correctly. In fact, it has been suggested that adaptive tests may be more motivating for low-ability students because they will not have to see the more difficult items that would appear on a traditional paper-and-pencil test (Bergstrom et. al., 1992).

When taking a CAT, one potential problem for low-ability students concerns the issue of item difficulty. Traditional paper-and-pencil exams are generally

written so item difficulties are arranged in ascending order (Hambleton & Traub, 1974; Plake, Ansorge, Parker, & Lowrey, 1982; Allison, 1984; Tippetts & Benson, 1989; Ackerman, Spray, Reckase, & Carlson, 1989). This is particularly beneficial for a low-ability student when taking a timed test. If they are unable to complete the test, they will not have time to attempt the more difficult items. This would not be considered a large problem because such students would probably answer many, if not most, of these items incorrectly. Typically on a CAT test, items of medium difficulty are presented first. As these items are answered incorrectly the computer is programmed to administer successively easier items. A few studies on item order found that tests presented with a descending order of item difficulty produced significantly more stressful test situations than tests with an ascending order of item difficulty (Hambleton & Traub, 1974; Plake et al., 1982; Tippetts & Benson, 1989). Legg and Buhr (1992) suggested that increased anxiety on CATs may cause scores to be lower than on conventional paper-and-pencil tests.

When addressing the issue of CAT items being unduly challenging there seems to be one basic question. That is, when considering the psychometric consequences of administering items that are slightly easier than an individual's current ability estimate, one is asking how statistically robust is the procedure when one deviates from administering ideal items, to a particular degree. It is also important to consider what constitutes a "slightly" easier item. That is, how much easier may items become before noticing a major impact on the number of items that need to be administered during a test, or on the actual computation of

θ , if the stopping criterion is a predetermined s.e.(θ). Additionally, if the stopping rule is based on number of items administered, what effect will there be on the s.e.(θ)? Bergstrom, et. al. (1992) and Lunz and Bergstrom (1994) investigated similar questions by administering mastery test items to three different groups of subjects where each group had a 50%, 60% and 70% chance of answering items correctly, respectively. Both of these studies employed a stopping rule where the estimated score was within a pre-determined distance from the decision point. Both studies found no significant differences with respect to estimated ability. With a 70% chance of correct response, the results of both studies indicated that there was some increase in test length to reach the desired level of accuracy, but noted that the practical impact was minimal.

Bergstrom et. al. (1992) suggest that, when considering the fact that higher ability students are used to answering well over 50% of the questions on a test correctly, test developers may choose to select an algorithm that will administer easier test items. This may have the effect of reducing anxiety and giving students an edge so they have the chance of answering a higher percentage of test items correctly. These researchers realize that this may result in a slightly longer test, but they think the potential psychological impact of making examinees take very hard tests may not be worth the savings in test length. This issue of increased test length seems even less important when considering that CATs generally contain significantly fewer items than traditional paper-and-pencil tests (McBride, 1985; McBride & Martin, 1983; Plake, 1995; Straetmans & Eggen, 1998; Wainer, 1990; Weiss, 1985). Lunz and Bergstrom (1994) set out

to explore the issues that would be of most concern to those test takers who will challenge computerized adaptive tests, and they believe that the most able students will probably feel that CATs are more difficult than paper-and-pencil tests. They state that we must have valid measurements but those which provide a reasonable level of comfort for the examinees.

Ponsoda, Olea, Rodriguez and Revuelta (1999) conducted related research to determine the effects of manipulating test difficulty on ability estimates and anxiety levels using both CATs and self-adaptive tests (SATs). Unfortunately in this research only easy and difficult conditions were compared. No baseline comparison condition was employed. This condition would have been one in which targeted item difficulty is equal to the current ability estimate, according to IRT. In their easy condition, targeted item difficulty = current ability estimate – 0.5. In the difficult condition, targeted item difficulty = current ability estimate + 0.5 (Ponsoda, et. al., 1999). In terms of how these manipulations affect one's probability of answering any particular item correctly, the easy condition gives an individual a 70% chance of a correct response. The difficult condition gives an individual only a 30% chance of a correct response. Here a Rasch model was employed with a scaling constant of $D = 1.7$.

One may wonder why Ponsoda et. al. (1999) did not compare the easy and difficult conditions to the optimal condition in which one has a 50% chance of correctly responding to any particular item. Additionally, one may wonder why one would consider administering items that are so difficult that an individual only has a 30% chance of correctly responding to those items, especially when this

study considered pre- and post-test anxiety levels. It would seem more beneficial to compare the optimal condition, where $P(\theta) = 0.5$, to an easier condition in order to see how ability estimates and anxiety levels are effected. It is no surprise to see this research resulting in significant differences in the pre-test – post-test anxiety change scores when comparing the easy and difficult conditions. One would hope that this anxiety change score would have also resulted in significant differences between easy and optimal conditions, had this optimal condition been considered. These results would provide psychological evidence for making CAT tests easier than the optimal condition allows.

Chapter III

METHOD

A. Rationale for Research

Bergstrom et. al. (1992) and Lunz and Bergstrom (1994) conducted studies that employed CATs programmed to administer items at the optimal level (50% chance of a correct answer) and two distinct sets of easier items, those in which individuals had a 60% or 70% chance of correctly responding to items presented. These studies used a stopping rule that was based on precision in making pass/fail decisions on mastery tests. Both of these studies investigated the effect of altering item difficulty on pass/fail decisions, which were based on mastery test scores.

Ponsoda, et.al. (1999) looked at the effect of item difficulty manipulation on ability estimation and pre-test – post-test anxiety differences. In addition, this study considered a fixed number of test items ($n = 20$) and looked at the effects of two conditions on s.e. (θ): easy, where the probability of correctly answering an item is 0.7, and difficult, where the probability of correctly answering an item is 0.3. As previously stated, it is puzzling why these researchers did not consider the optimal condition, psychometrically speaking, where an individual has a probability of answering an item correctly of 0.5.

B. Overview of Procedure

This study blends the work of Bergstrom et. al. (1992), Lunz and Bergstrom (1994), and Ponsoda et. al. (1999), investigating the effects of manipulating item difficulty on (a) the number of items needed to achieve a predetermined level of precision across the continuum of ability scores and (b) ability estimation. In the previously mentioned studies, only two changes were made with respect to item difficulty and only one stopping rule was investigated.

This study considers eleven levels of true ability, ranging from -2.5 to $+2.5$. Also considered were eight different levels of manipulation concerning item difficulty: $\delta = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8$, and 1.0 , that is items were made δ units easier than optimal items. In addition, three pre-determined levels of precision for the stopping rule were considered: $s.e.(\theta)$ of $0.2, 0.35$, and 0.5 . Using a $s.e.(\theta)$ of 0.2 attempted to achieve very low measurement error. This corresponds to a reliability of 0.96 using the previously discussed relationship described by Urry (1977). Using a $s.e.(\theta)$ of 0.35 corresponds to a reliability of 0.88 , and a $s.e.(\theta)$ of 0.5 corresponds to a reliability of 0.75 , which is about the same as, or slightly higher than, a standard teacher-made test. Finally, different numbers of initial items (three, five or seven) were administered prior to the start of the adaptive portion of the test. These investigations were conducted through the use of computer simulations.

Initially the standard IRT approach of matching b (item difficulty) to $\hat{\theta}$ (current ability estimate) to determine the difficulty level for subsequent items

was employed. Each level of modification to item difficulty, number of initial items, and level of precision was applied across ability levels throughout the study which makes this a $11 \times 9 \times 3 \times 3$ factorial design:

$$[\text{true } \theta] \times [\delta] \times [\text{s.e.}(\theta)] \times [\text{initial items}]$$

Table 1 indicates how $P(\theta)$, the probability of correctly answering an item, is affected by letting $b = \theta - \delta$. With the standard IRT approach, using optimal items ($\delta = 0$) results in $P(\theta) = 0.5$. That is, when item difficulty exactly matches an individual's ability level, that individual's chances of answering that item correctly are 50%. As δ increases, and item difficulty therefore decreases, one's chances of answering an item correctly increase. For example, when an item is 0.1 units easier on the item difficulty scale, an individual's chances of answering that item correctly are increased to 54.2%. If an item is 1 full unit easier, an individual's chances of a correct response increase to 84.6%.

Table 1: The effect of δ on $P(\theta)$

δ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
$P(\theta)$	0.5	0.542	0.584	0.625	0.664	0.701	0.735	0.796	0.846

δ = the modification made to item difficulty

$P(\theta)$ = probability of answering an item correctly

Recall that when $b = \theta$ the standard error of the estimated value for θ ($\hat{\theta}$), $\text{s.e.}(\hat{\theta})$, is minimized. Thus, item information, $I(\theta) = 1/[\text{s.e.}(\hat{\theta})]^2$, is maximized. Item information cumulatively contributes to test information. The larger the test information, the more precise the test. As items became easier, with each

successive version of the program, it was expected that additional items would need to be administered in order to achieve the specified level of precision. That is, as δ increased and test items became easier and easier, more and more items would be required. This is because each of the easier items contributed less information than optimal items. So more items were needed to accumulate the necessary test information corresponding to the desired level of precision. Because the parameters in item response models are invariant, it was also expected that there would be no noticeable impact of item difficulty manipulation on the estimation of ability.

C. Theoretical Examination

Theoretically speaking, it is possible to determine the minimum number of items necessary to reach the specified levels of precision given an ideal testing situation. That is to say if one were working under ideal circumstances and had precise information about an examinee's true ability level, a test could be made to consist of optimal items. These items would have the precise difficulty level to exactly match that examinee's true ability level, θ . This, of course, is impossible. One can never fully know an examinee's true ability because it is a latent variable. One can only estimate true ability based on observed test scores. In any case, if such a test could be constructed and all items had the same difficulty level, $b = \theta$, each item would contribute the same amount of information to the total test information. Since test information is summed over

all items, the total test information would be equivalent to the number of test items multiplied by the amount of item information contributed by each test item [$I(\theta) = n I_i(\theta)$].

Using the 1 – parameter logistic model, $P(\theta) = \frac{1}{1 + e^{-1.7(\theta - b_i)}}$ and given that

$$I(\theta) = \frac{[P'(\theta)]^2}{P(\theta)Q(\theta)}, \text{ where } Q(\theta) = 1 - P(\theta), \text{ for this model } I(\theta) = \frac{1.7^2 e^{-1.7(\theta - b_i)}}{[1 + e^{-1.7(\theta - b_i)}]^2}.$$

When $b_i = \theta$, $I(\theta) = 0.7225$. Since $I(\theta) = \frac{1}{[s.e.(\theta)]^2}$, one can also show that the

number of items required to reach a specified s.e.(θ) is $n = \frac{1}{[s.e.(\theta)]^2 I_i(\theta)}$. Table

2 shows the number of items, n , required to reach each of the three s.e.(θ) values that were considered in this study. Table 2 also indicates the increase in n over what is required when optimal items are employed (when $\delta = 0$) with the corresponding percentage increase. Recall that this table only provides $I(\theta)$ and n in the ideal situation where all b_i are exactly the same. This would never actually occur in any testing situation, real or simulated.

From the data in Table 2 it can be seen that when δ is 0.1 or 0.2 there is a negligible increase in n for all s.e.(θ). The increase in all cases is approximately one to four percent. When δ is 0.3 the increase is approximately 6 to 7%. This increase is less than one item when more measurement error is permitted. When s.e.(θ) is 0.2 the increase is only 2.3 items. But it is important to note that practical significance is of importance here. When δ is 0.4 and n is increased by 12 to 13%, even in the most restrictive situation where s.e.(θ) is 0.2, the test

Table 2: Item Information, number of test items and corresponding increase over optimal CAT condition (% increase) given modification to item difficulty.

δ	$I(\theta)$	n when s.e. (θ) is			increase in n (%) when s.e. (θ) is		
		0.2	0.35	0.5	0.2	0.35	0.5
0	0.7225	34.6	11.3	5.5	0	0	0
0.1	0.7173	34.9	11.4	5.6	0.3 (0.9)	0.1 (0.9)	0.1 (1.8)
0.2	0.7020	35.6	11.6	5.7	1.0 (2.9)	0.3 (2.7)	0.2 (3.6)
0.3	0.6775	36.9	12.0	5.9	2.3 (6.6)	0.7 (6.2)	0.4 (7.3)
0.4	0.6450	38.8	12.7	6.2	4.2 (12.1)	1.4 (12.4)	0.7 (12.7)
0.5	0.6062	41.2	13.5	6.6	6.6 (19.1)	2.2 (19.5)	1.1 (20.0)
0.6	0.5629	44.4	14.5	7.1	9.8 (28.3)	3.2 (28.3)	1.6 (29.1)
0.8	0.4697	53.2	17.4	8.5	18.6 (53.8)	6.1 (54.0)	3.0 (54.5)
1.0	0.3775	66.2	21.6	10.6	31.6 (91.3)	10.3 (91.2)	5.1 (92.7)

$I(\theta)$ = test information

δ = the modification made to item difficulty

n = number of test items

is only increased in length by approximately four items. In the other two situations where more measurement error is permitted, there is only approximately one additional item that needs to be added to the test. One can see that as δ increases, the number of additional test items becomes quite large when s.e. (θ) is 0.2, but for the less precise situations the increase in n may still be acceptable to many test designers. It would be difficult to argue that lengthening a test by two or three items (when δ is 0.6 and s.e. (θ) is 0.35 or 0.5) would be of great concern, especially considering the potential psychological

impact of taking a test which is composed of much easier items yet still yields the same accurate results. Recall that when δ is 0.6 one's chances of a correct response increase to 73.5%.

The data from Table 2 provide an ideal set of results that would be expected after making the modifications in item difficulty as discussed in this study. In the actual study, similar results were expected for ability levels near the middle of the ability continuum where θ is close to zero. This is because the first items to be administered were of average difficulty ($b_i = 0$). Minor increases in test length over the ideal situation were considered possible because in the study three, five, or seven items were administered before an initial estimate of ability was determined. (From the fourth, sixth, or eighth item on, ability was re-estimated after each new item was administered.) As scores shifted away from the middle of the ability continuum (above or below), an additional increase in test length, over the ideal situation described above, was expected because the earlier items on the test would not provide as much information as optimal items. The farther out on the ability continuum, the less information about ability was provided.

D. Participants

This simulation was run for 1100 hypothetical individuals, 100 at each value of θ ranging from -2.5 to 2.5 , increasing by increments of 0.5 .

E. CAT Simulation

A computer program, written in Qbasic, simulated a CAT exam for the 1100 hypothetical individuals. The computer program to implement this is found in Appendix A. A Birnbaum equivalent of the Rasch model was employed for these simulations. A scaling constant of 1.7 was used. Recall that when $D = 1.7$ there is less than a 0.01 difference between the logistic model and the normal ogive (Birnbaum, 1968; Hambleton & Swaminathan, 1985). Using this model, it is assumed that all item discrimination scores are the same and that all items are not subject to random guessing, so that $c = 0$ for all items. The model is

$$P(\theta) = \frac{1}{1 + e^{-1.7(\theta - b_i)}}$$

The first step was to choose a true θ value corresponding to an individual from the ability range $\theta \in [-2.5, 2.5]$. The second step was to administer a given number of initial CAT items, in this case three, five, or seven, respectively, in each major version of the program. Each initial item administered was of average difficulty, $b_1 = b_2 = b_3 = 0$. The responses to these initial items provided the data from which an initial estimate for θ was computed. The initial items were all of average difficulty because in actual testing situations there is generally no prior knowledge about an examinee's ability level.

The third step was to estimate ability, $\hat{\theta}$, so the test follows an adaptive algorithm of administering items that match the current estimate of θ (or, in this study, a modified value of $\hat{\theta}$). To estimate θ after the first three items had been

administered, the closed form estimate, $\hat{\theta} = \frac{\log \frac{r}{n-r}}{1.7}$, was calculated for any

individual who answered at least one question correctly and one question

incorrectly. Here r = number of correctly answered items and n = total number of

items administered. If any individual answered all questions either correctly or

incorrectly, the above estimate for θ is undefined. In either of these two cases

the following Bayesian adjustment to the estimate was made: $\hat{\theta} = \frac{\log \frac{r+0.5}{n-r+0.5}}{1.7}$.

A fourth, sixth, or eighth CAT item was then administered where $b_4 = \hat{\theta}_3 - \delta$,

$\delta \in [0, 1.0]$, when three initial items were used.

Random numbers were generated to simulate an individual's score, I_j , on any given item. For example, for the initial items administered, $P(I_j = 1 | \theta, b_j = 0) = 0.5$ (for $j = 1, 2, 3$ and up to $j = 7$, depending on the number of initial test items administered) when true ability is $\theta = 0$ (and different for other values of θ). A random number, $u_i \sim U(0, 1)$, was generated for each item. For all items in general, if $u_i \leq P(I_j = 1 | \theta, b_j = \hat{\theta}_j - \delta)$, it was assumed that $I_j = 1$, that is, the item was answered correctly, where $\hat{\theta}_j$ is the current ability estimate. After each new item was administered, from the fourth, sixth, or eighth item on, θ was estimated and $s.e.(\hat{\theta})$ was computed.

During a CAT the computer is continuously re-estimating an individual's ability level based on item difficulties and examinee responses. In order to estimate θ from the fourth, sixth, or eighth item to the end of the test, maximum likelihood methods were employed, such as the iterative Newton-Raphson

numerical procedure (Hambleton & Swaminathan, 1985). Maximum likelihood estimation is the preferred method by many because maximum likelihood estimates (MLE) are consistent, that is, as sample size and number of items increase, the estimated value for θ approaches the true value for θ (Hambleton & Swaminathan, 1985). Additionally, MLEs are asymptotically normal and efficient, i.e., they have the smallest variance (Birnbaum, 1968; Hambleton & Swaminathan, 1985).

In computing a MLE the likelihood function, L , is specified by the equation:

$$L = L(u_1, u_2, u_3, \dots, u_n | \theta) = \prod_1^n P_i^{u_i} (1 - P_i)^{1-u_i}. \text{ The MLE of } \theta \text{ provides the value of } \theta$$

which maximizes L , though the log likelihood function is often employed and produces the same answer. The log likelihood function is given by the equation:

$$\ln L(u | \theta) = \sum_1^n [u_i P_i + (1 - u_i)(1 - P_i)] \quad \text{This function is maximized when}$$

$d/d\theta [\ln L(u | \theta)] = 0$. The standard error of the estimate, which was computed after each additional item had been administered, is determined by the equation

$$\text{s.e.}(\hat{\theta}) = \frac{1}{\sqrt{1.7^2 \sum P_i (1 - P_i)}} \quad ((\text{Hambleton \& Swaminathan, 1985}).$$

One problem when employing MLEs occurs when there is a response pattern of all correct or all incorrect responses. In these two situations the estimated value for θ approaches $+\infty$ or $-\infty$, respectively. In either of these two situations, a Bayesian approach was be taken in order to determine the current estimate for θ (Hambleton & Swaminathan, 1985; Hambleton et. al., 1991; Weiss & Kingsbury, 1984). Individuals with $r = 0$ were assigned $r = 0.5$.

Individuals with $r = n$ were assigned $r = n - 0.5$. This adjusted r value was then input into the iterative procedure to allow it to converge to a $\hat{\theta}$ value, which was used to determine the difficulty of the next item to be administered. The r value was corrected before proceeding on to the next item on the test. The stopping procedure employed here involved administering a minimum number of items (n) to each individual examinee in order to achieve the pre-determined level of precision: $\text{s.e.}(\hat{\theta}) < 0.2, 0.35, \text{ or } 0.5$.

Ordinarily, in a CAT situation, content balancing must be considered in order to be sure all required topics are covered. This may require a complicated item selection routine to insure that all content areas have been adequately represented. For the sake of simplicity, it was assumed here that there is only one content domain from which to select items. That is, no other item selection routine other than choosing items based on the current ability estimate, and the current modification to item difficulty considered in that version of the program, was considered here.

F. Analysis

For each version of the program, employing the three different numbers of initial test items (three, five and seven) and the three $\text{s.e.}(\theta)$ values of 0.2, 0.35 and 0.5, the average θ estimate was calculated for each true θ value at each δ value. That is, for the eleven true ability levels under consideration, the simulated CATs were "administered" and ability estimates were determined

under several conditions. First, ability estimates were determined when optimal CAT items are employed under a stopping rule of $\text{s.e.}(\hat{\theta}) < 0.2$, while also applying different numbers of initial test items. Ability estimates were determined for each successive modification concerning item difficulty. The decrease in item difficulty is indicated by δ . That is, easier items were chosen by 0.1 units, initially, until $\delta = 0.6$, and then by 0.2 units, until $\delta = 1.0$. An average of the θ estimates and the corresponding standard deviation were calculated for each δ value at each true θ value. The average number of items administered at each true θ value, and the corresponding standard deviation, were calculated for each δ value. Additional simulations were run to examine the above measures when employing stopping rules of $\text{s.e.}(\hat{\theta}) < 0.35$ and $\text{s.e.}(\hat{\theta}) < 0.5$.

The number of items required to reach the specified $\text{s.e.}(\hat{\theta})$ values were examined to determine if there was any sizeable increase in test length. The data were examined to determine the impact on test length, if any, as a result of administering different numbers of initial test items prior to the start of the adaptive portion of the test. Estimated values for θ were examined to determine if they remained relatively consistent across δ values. Again, it is necessary to realize that matters of practical importance were of concern here. It is possible to compare CATs where optimal items have been administered throughout the test to CATs where non-optimal items have been administered, as discussed above, and see statistically significant differences between ability scores resulting from tests that may actually differ by one or two test items.

Chapter IV

RESULTS

For this study a series of computer simulations were conducted to determine the effect of the manipulation of item difficulty on test length in a CAT environment. Also under examination is the effect of administering different numbers of initial test items, prior to the start of adaptive portion of the test, on test length. Finally, the effect, if any, on the consistency of ability estimates with respect to the two previous modifications is under consideration.

A. Manipulation of item difficulty

The data presented in Figures 6 – 14 (see Appendix B for data) illustrate the fact that the average number of test items is lowest for examinees of average ability ($\theta = 0$), as expected. The average number of test items increases, with few exceptions, as ability levels approach the extremes of the ability continuum. This pattern holds regardless of the level of precision under consideration or the number of initial test items administered prior to the start of the adaptive portion of the test.

The data illustrated in Figures 6 - 14 also make apparent the fact that as δ increases, i.e., as test items become easier, the number of test items needed to reach the desired level of precision, $s.e.(\theta)$, also increases, as expected. However, in general, as $s.e.(\theta)$ increases and the results of these tests become less precise, fewer extra test items are required.

Figure 6: Average number of test items (n) for 3 initial test items and s.e.(θ) of 0.2

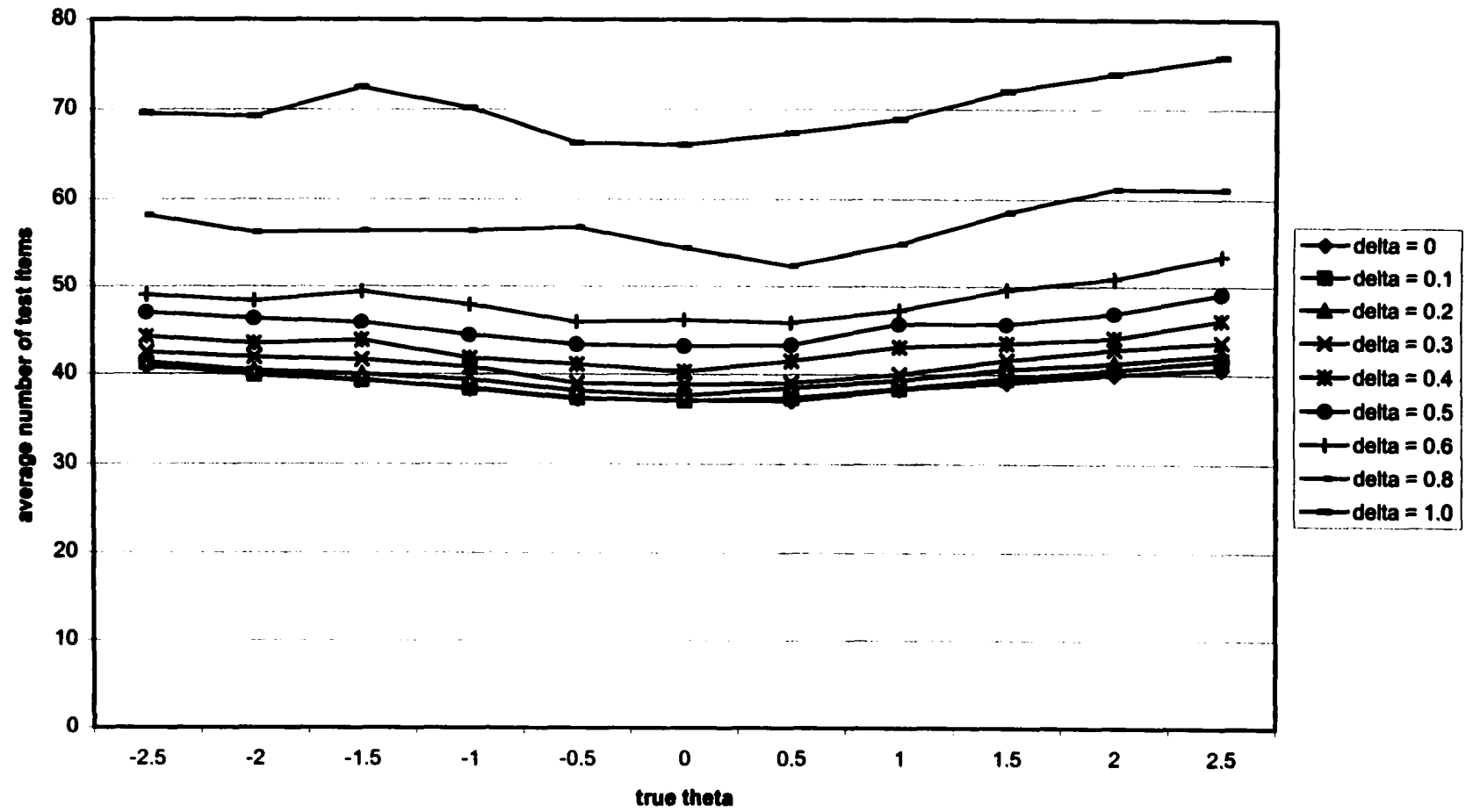


**Figure 7: Average number of test items (n) with 5 Initial test items
and s.e.(theta) of 0.2**

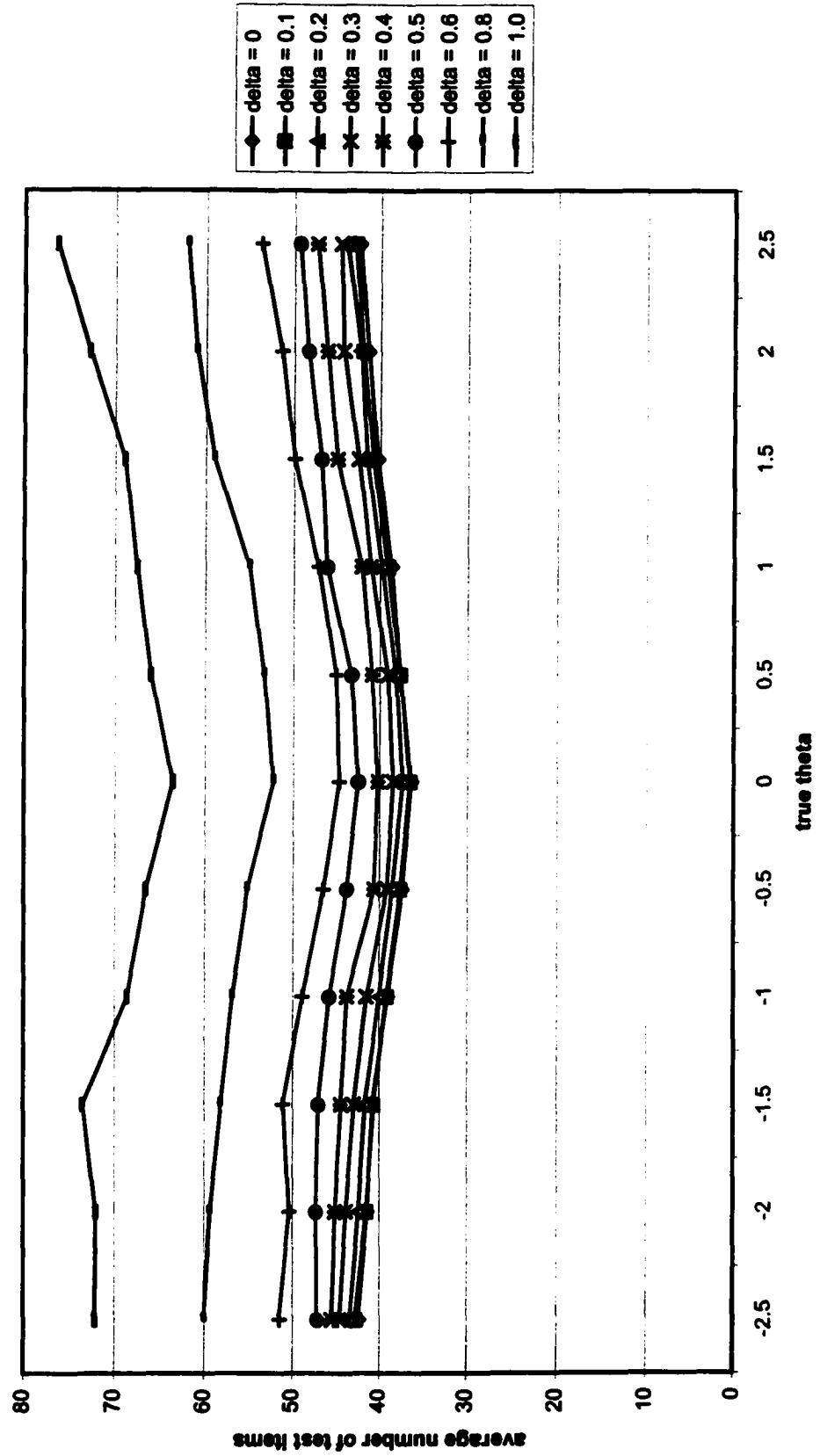


Figure 8: Average number of test items (n) with 7 initial test items and s.e.(theta) of 0.2

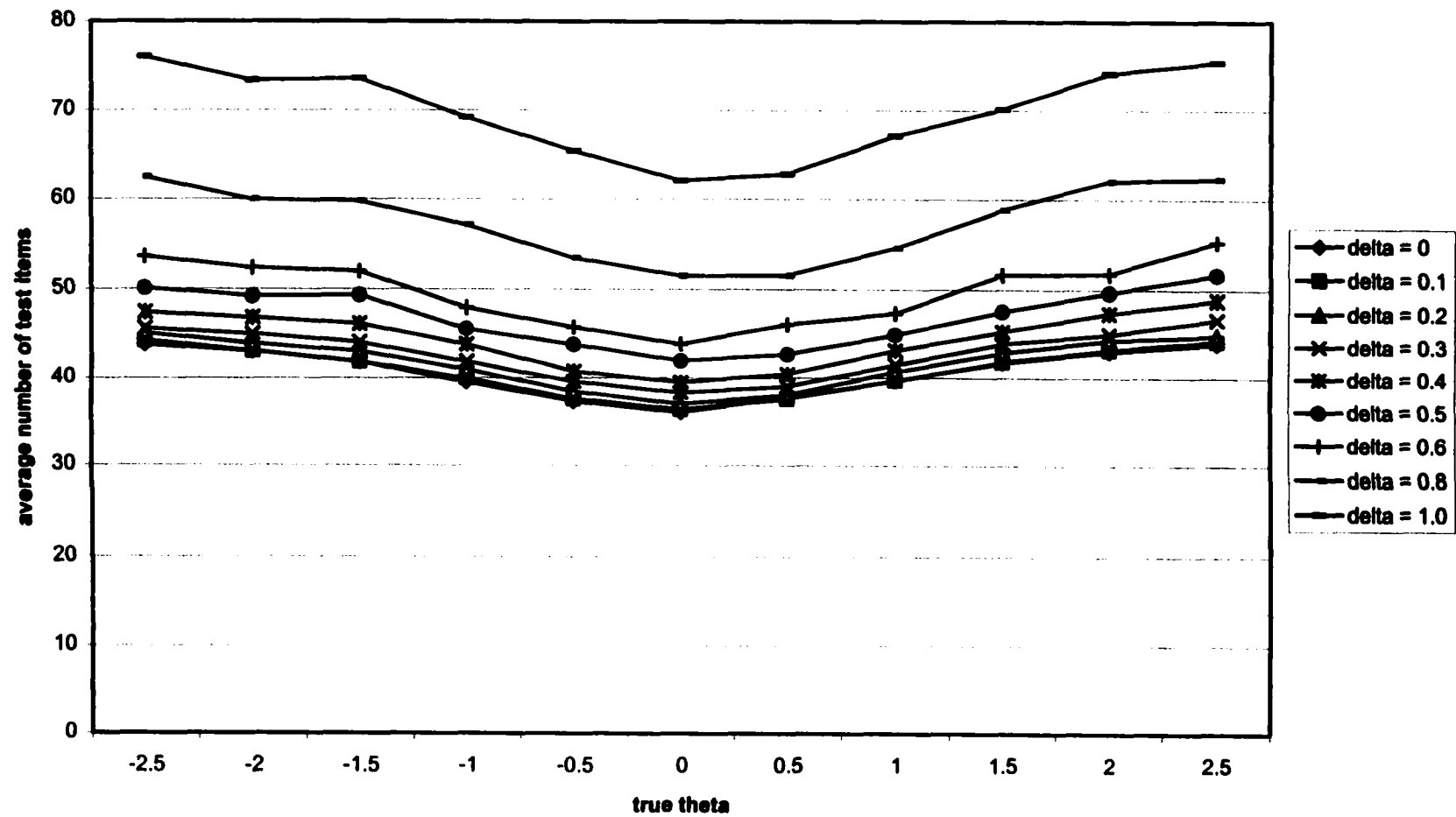


**Figure 9: Average number of test items (n) with 3 Initial test items
and s.e.(theta) of 0.35**

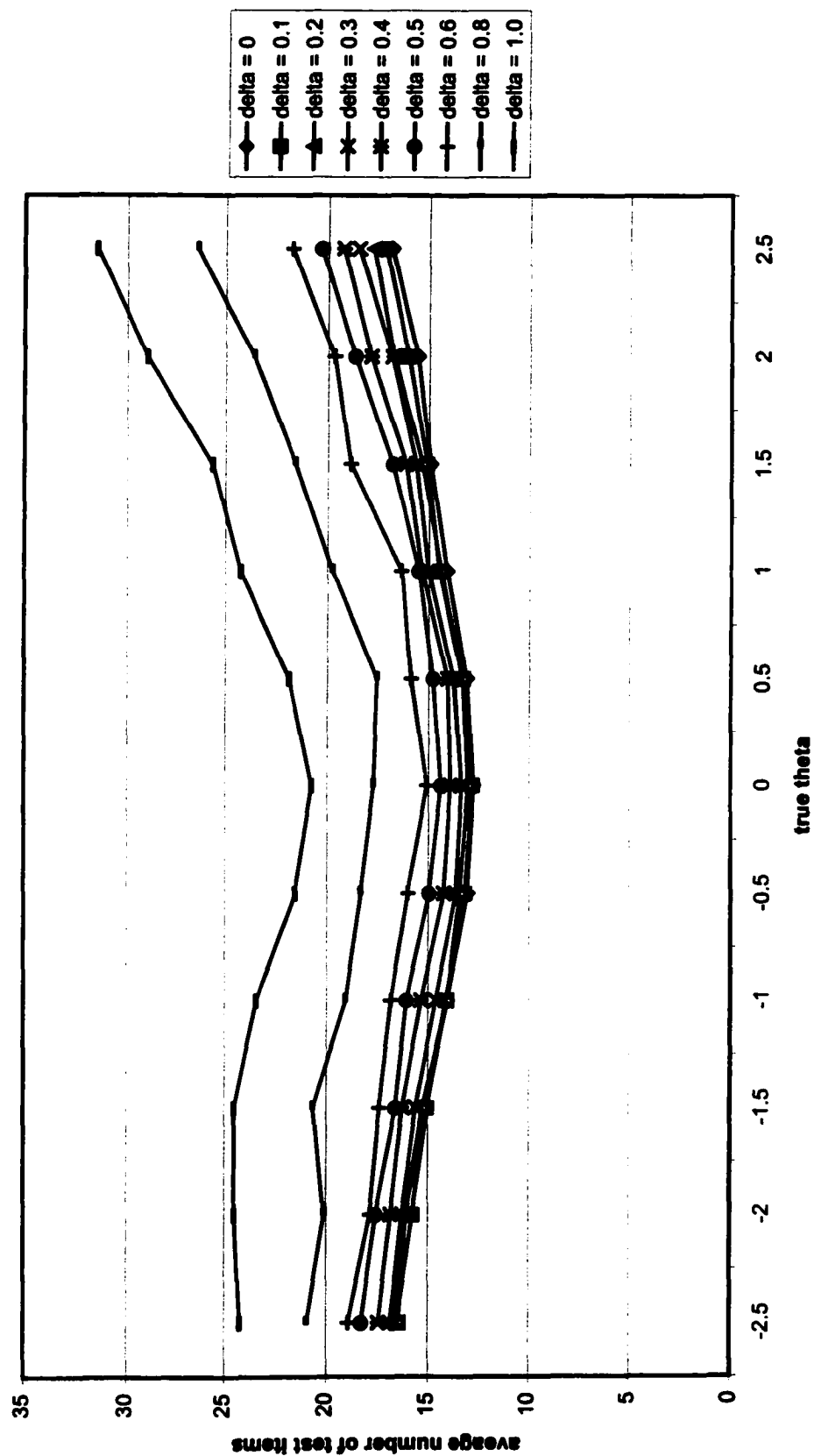


**Figure 10: Average number of test items (n) with 5 initial test items
and s.e.(theta) of 0.35**

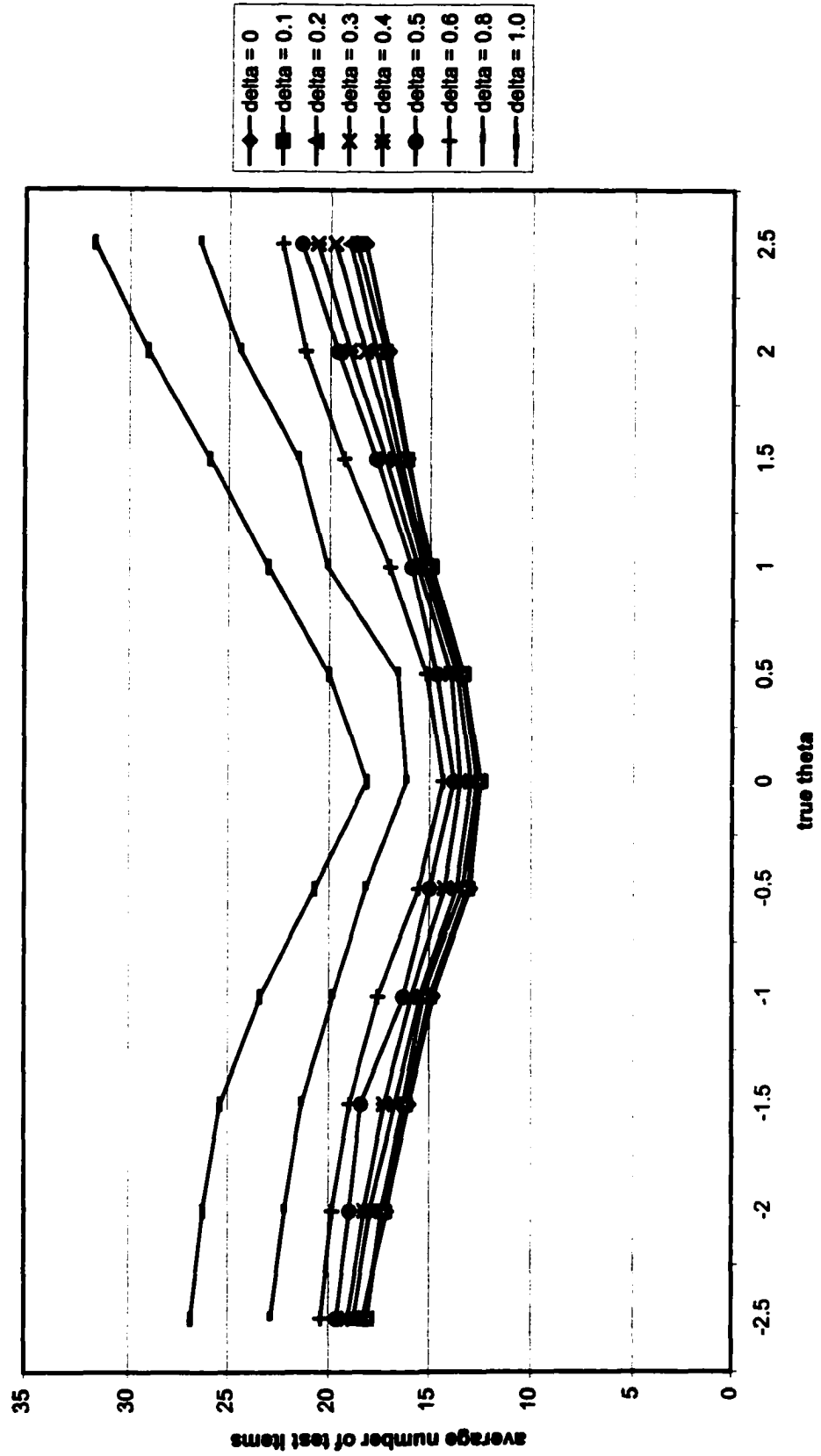


**Figure 11: Average number of test items (n) with 7 initial test items
and s.e.(theta) of 0.35**

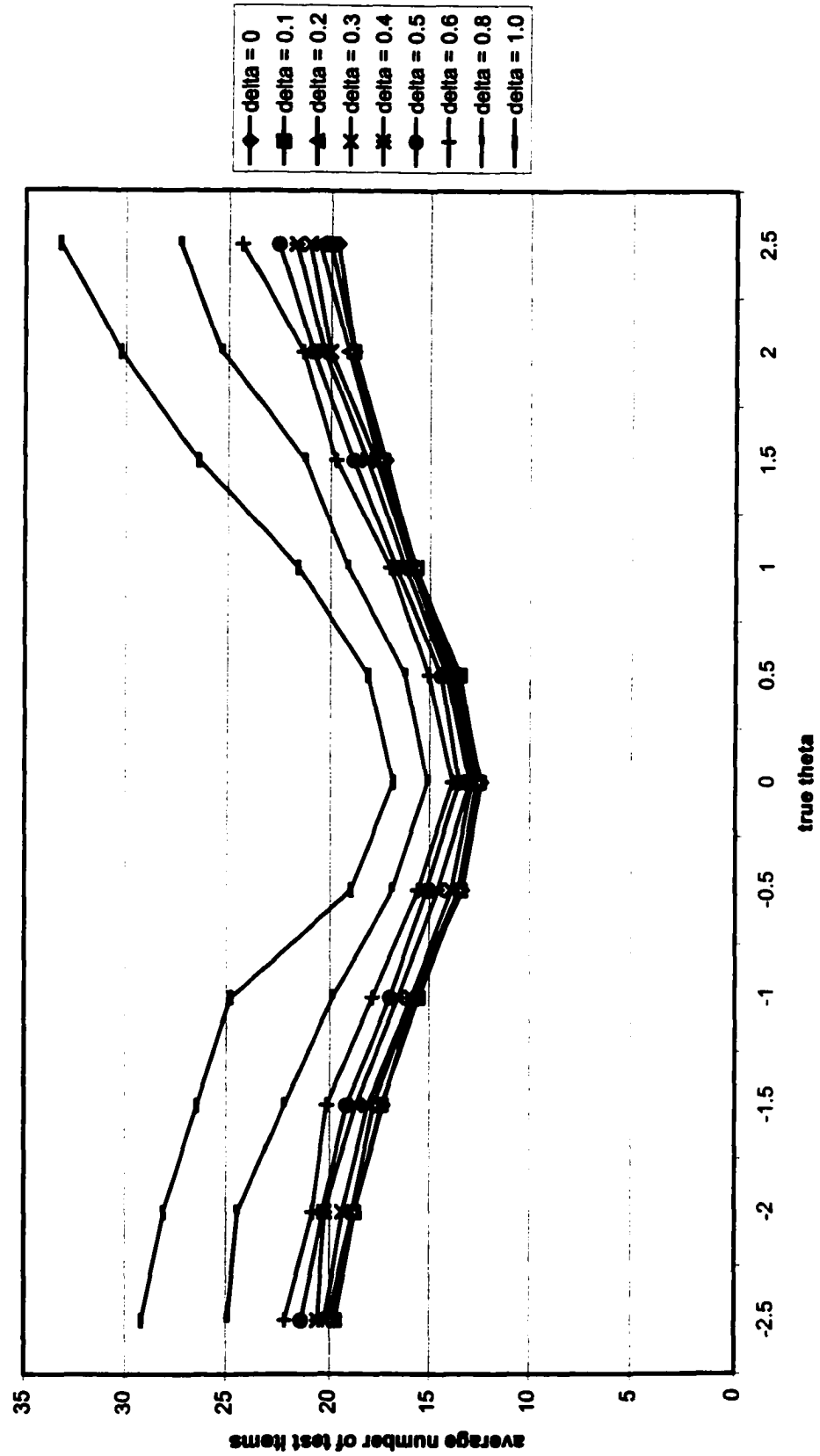


Figure 12: Average number of test items (n) for 3 Initial test items and s.e.(theta) of 0.5

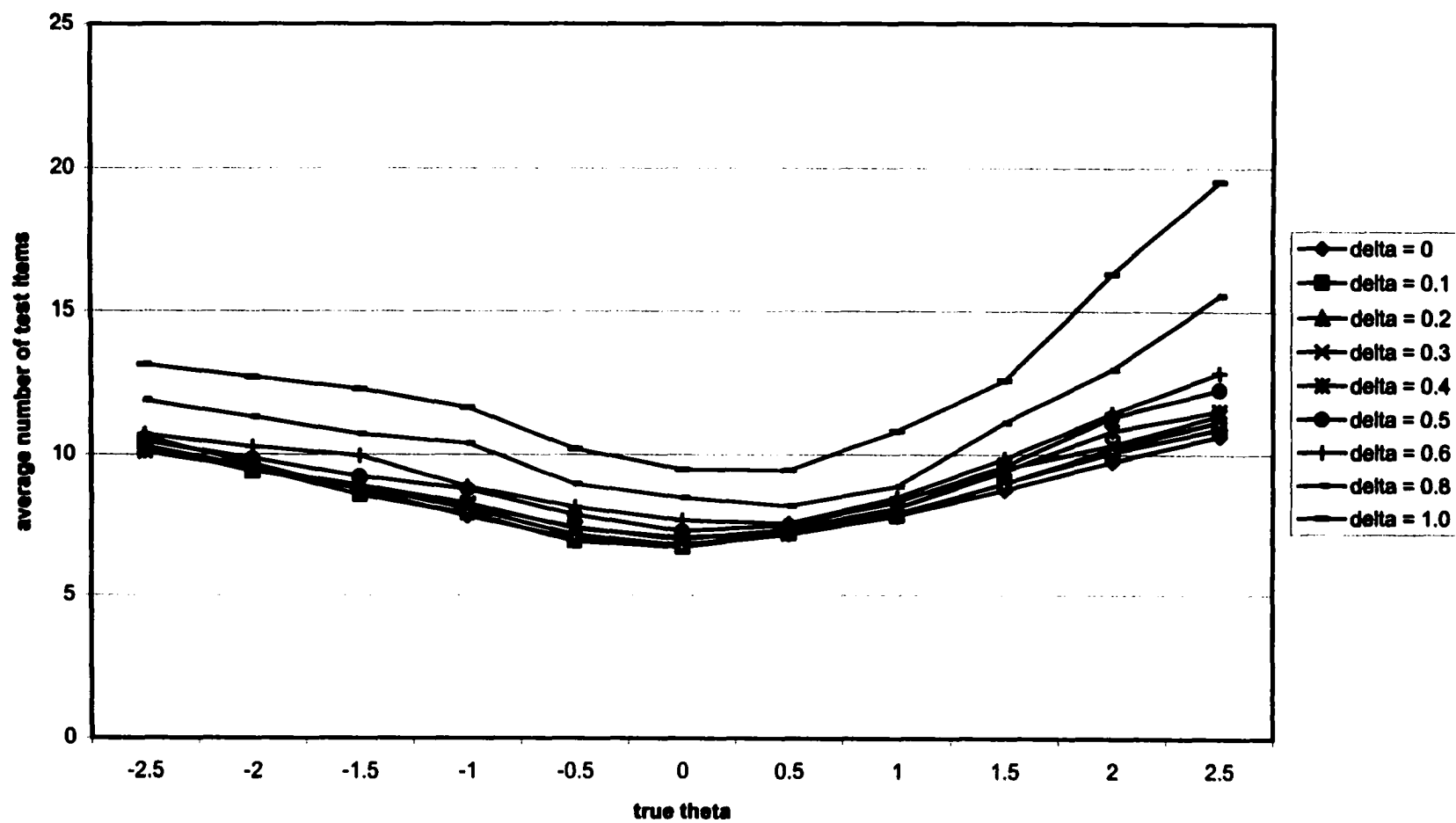


Figure 13: Average number of test items (n) with 5 initial test items and s.e.(theta) of 0.5

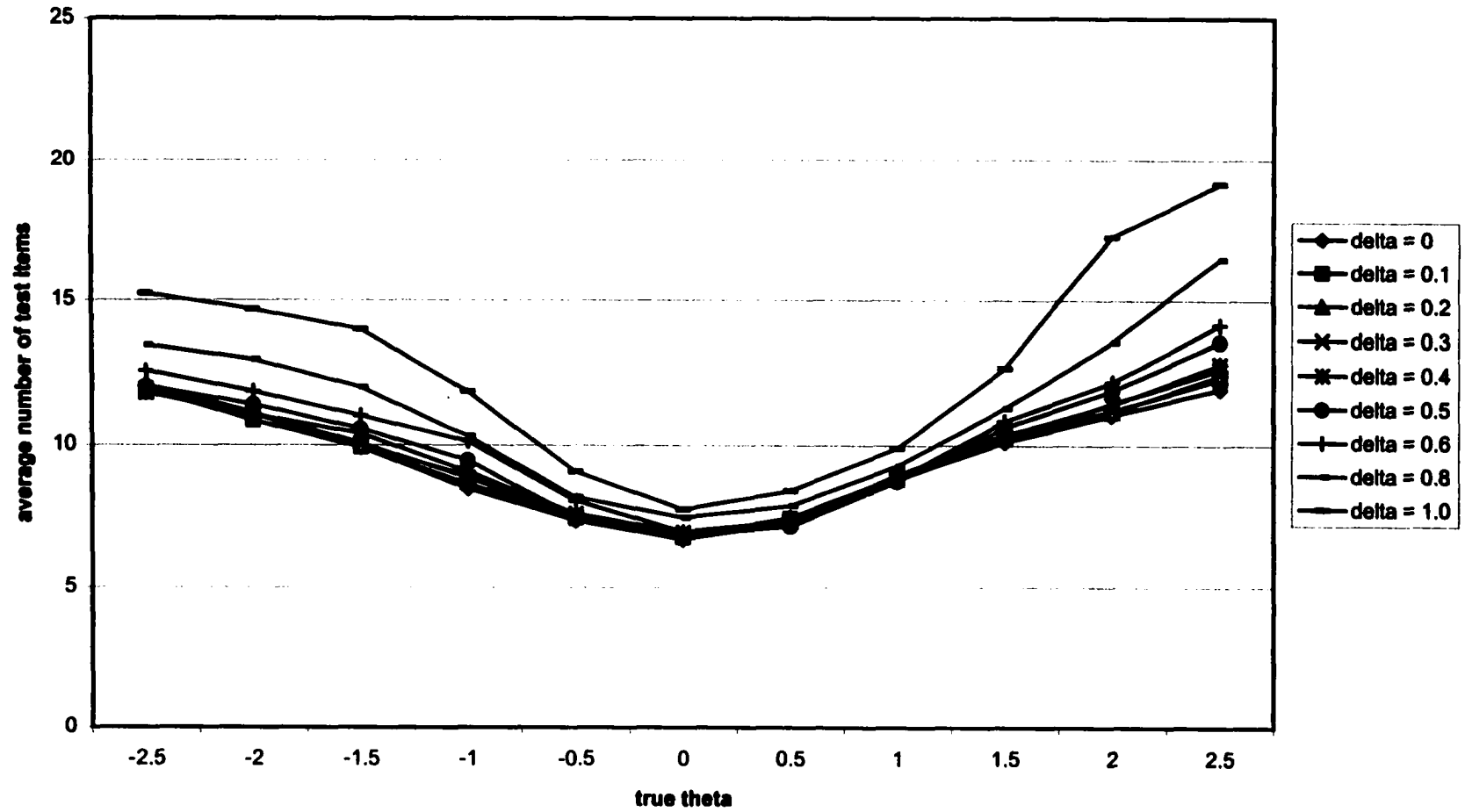
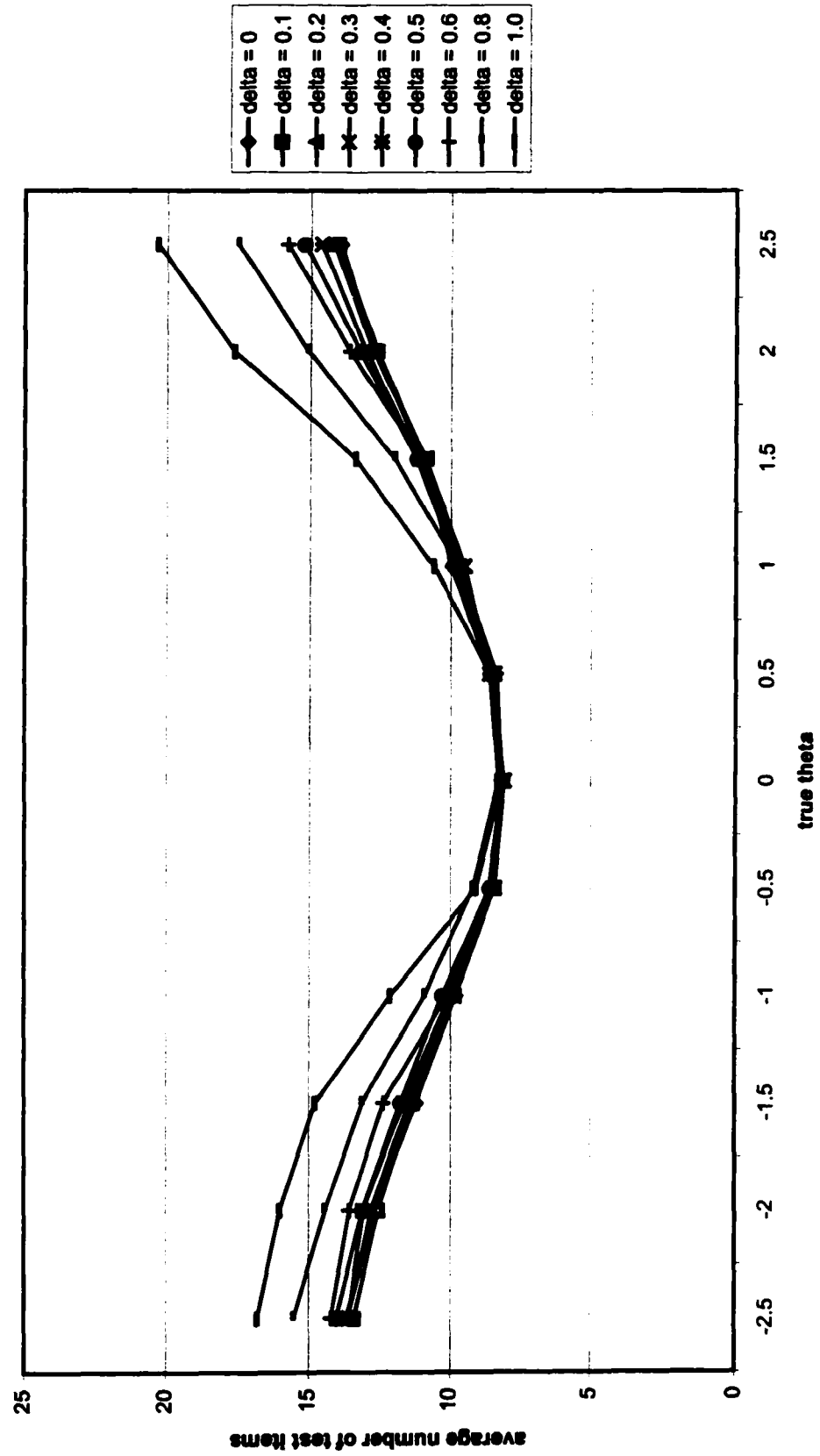


Figure 14: Average number of test items (n) with 7 initial test items and s.e.(θ) of 0.5



Tables 3, 4, and 5 provide a summary of the data presented in Appendix B. The data are averaged across ability levels to show the average number of test items required to reach the desired levels of precision. Corresponding increases in test length over the optimal condition of $b = \theta$ ($\delta = 0$) are provided along with the percentage increase represented. When considering the effect of the manipulation of item difficulty on the number of test items, there are many δ values that only cause a slight increase in test length.

When averaging over all ability levels, under a number of conditions, certain δ values only increase test length by approximately one item or less. For example, when $\text{s.e.}(\hat{\theta}) < 0.2$, this modest increase occurs for $\delta = 0.1$ and 0.2 , representing up to a 3% increase in test length. When $\text{s.e.}(\hat{\theta}) < 0.35$ these results occur for $\delta = 0.1$ to 0.4 . The largest increase occurring here represents an 8.1% increase in test length. When $\text{s.e.}(\hat{\theta}) < 0.5$ this modest increase occurs for $\delta = 0.1$ to 0.6 . All of the above mentioned results occur regardless of the number of initial test items administered prior to the start of the adaptive portion of the test. This modest increase also occurs when seven initial items are administered and for $\text{s.e.}(\hat{\theta}) < 0.5$ and $\delta = 0$.

Increases of approximately two to four test items occur when $\text{s.e.}(\hat{\theta}) < 0.2$ for $\delta = 0.3$ and 0.4 . Here increases in test length of 5.1% to 10.4% are represented. When $\text{s.e.}(\hat{\theta}) < 0.35$, similar increases occur for $\delta = 0.5$ and 0.6 . Here only approximately two to three additional test items are needed which represent increases in test length of 9.4% to 19.1%.

Table 3: Average number of test items and percent increase over optimal CAT items with s.e.(θ) of 0.2.

Delta	3 initial items			5 initial items			7 initial items		
	Average n	Average increase	% increase	Average n	Average increase	% increase	Average n	Average increase	% increase
0	38.93	-	-	39.83	-	-	40.71	-	-
0.1	39.15	0.22	0.6%	40.07	0.24	0.6%	40.88	0.17	0.4%
0.2	39.89	0.96	2.5%	40.83	1.00	2.5%	41.94	1.23	3.0%
0.3	41.05	2.12	5.4%	42.00	2.17	5.4%	42.80	2.09	5.1%
0.4	42.96	4.03	10.4%	43.75	3.92	9.8%	44.54	3.83	9.4%
0.5	45.49	6.56	16.9%	46.15	6.32	15.9%	46.98	6.27	15.4%
0.6	48.50	9.57	24.6%	49.11	9.28	23.3%	49.63	8.92	21.9%
0.8	56.93	18.00	46.2%	57.47	17.64	44.3%	57.59	16.88	41.5%
1.0	70.14	31.21	80.2%	69.80	29.97	75.2%	69.92	29.21	71.8%

Test items are averaged over subjects and ability levels.

Table 4: Average number of test items and % increase over optimal CAT items ($\delta = 0$, $b = \theta$) with $s.e.(\theta) < 0.35$.

Delta	3 initial items			5 initial items			7 initial items		
	Average n	Average increase	% increase	Average n	Average increase	% increase	Average n	Average increase	% increase
0	14.75	-	-	15.59	-	-	16.60	-	-
0.1	14.86	0.11	0.7%	15.68	0.09	0.6%	16.63	0.03	0.2%
0.2	15.07	0.32	2.2%	15.89	0.30	1.9%	16.87	0.27	1.6%
0.3	15.46	0.71	4.8%	16.28	0.69	4.4%	17.18	0.58	3.5%
0.4	16.06	1.31	8.9%	16.77	1.18	7.6%	17.66	1.06	6.4%
0.5	16.73	1.98	13.4%	17.39	1.80	11.5%	18.19	1.59	9.6%
0.6	17.70	2.95	20.0%	18.33	2.74	17.6%	18.91	2.31	13.9%
0.8	20.50	5.75	39.0%	20.87	5.28	33.9%	21.15	4.55	27.4%
1.0	24.65	9.90	67.1%	24.59	9.00	57.7%	24.89	8.29	49.9%

Test items are averaged over subjects and ability levels.

Table 5: Average number of test items and % increase over optimal CAT items ($\delta = 0$, $b = \theta$) with $\text{s.e.}(\theta) < 0.5$.

Delta	3 initial items			5 initial items			7 initial items		
	Average n	Average increase	% increase	Average n	Average increase	% increase	Average n	Average increase	% increase
0	8.62	-	-	9.54	-	-	10.93	-	-
0.1	8.62	0.01	0.1%	9.63	0.09	0.9	10.88	-0.05	-0.5%
0.2	8.73	0.11	1.3%	9.68	0.14	1.5%	10.93	-	-
0.3	8.89	0.27	3.1%	9.70	0.16	1.7%	10.94	0.01	0.1%
0.4	8.94	0.32	3.7%	9.80	0.26	2.7%	11.08	0.15	1.4%
0.5	9.32	0.70	8.1%	9.99	0.45	4.7%	11.26	0.33	3.0%
0.6	9.62	1.00	11.6%	10.37	0.83	8.7%	11.42	0.49	4.5%
0.8	10.78	2.16	25.1%	11.17	1.63	17.1%	12.22	1.29	11.8%
1.0	12.57	3.95	45.8%	12.72	3.18	33.3%	13.41	2.48	22.7%

Test items are averaged over subjects and ability levels.

All of the remaining conditions concerning the different number of initial test items administered and modifications made to item difficulty under the least precise condition, $\text{s.e.}(\hat{\theta}) < 0.5$, result in increases to test length of approximately two to four items. Here $\delta = 0.8$ and 1.0 , where one's chances of a correct response increase to 79.6% and 84.6%, respectively.

Slightly larger increases of six to seven additional test items occur for $\delta = 0.5$ under the most precise condition of $\text{s.e.}(\hat{\theta}) < 0.2$. The additional items represent increases of 15.4% to 16.9%. When $\text{s.e.}(\hat{\theta}) < 0.35$ approximately five to six additional test items result when $\delta = 0.8$. Increases in test length under this condition are 27.2% and 38% respectively.

Under all other conditions not previously discussed, resulting increases are approximately eight items or more. These represent increases ranging from 21.9% to 80.2%.

When comparing the theoretical data presented in Table 2 to the summary data, averaged over ability, in Tables 3, 4, and 5, one can see that, in all cases, the simulations result in tests that are just a few items longer than in the theoretical examination. This occurs regardless of the number of initial test items employed or the level of precision considered. When comparing the theoretical data presented in Table 2 to the summary data presented in Appendix B, one can see that the same results generally occur for each true θ value, averaged over subjects. It is interesting to note that for the highest levels of δ , when true θ is at or near average ability, average test lengths are actually shorter than the

test lengths resulting from the theoretical examination. This is likely due to the large variations in test length that occur at the higher δ values.

Upon examination of the data presented in Figures 6 through 14 (see Appendix B for data) another pattern emerges. The number of test items required to achieve the desired levels of precision, as tests increase in length, appear to be relatively symmetrical within the θ distribution at least for the lower δ values. As δ increases throughout the simulation, tests are increasing in length as well, as expected. For most of the conditions considered here, this symmetry appears to deteriorate as δ increases, i.e., as items become easier.

For low values of δ there is generally less than a one item difference in test length for ability levels at the extremes of the ability continuum, i.e., for $\theta = -2.5$ and $\theta = 2.5$. As δ increases this situation changes and the differences between test lengths for $\theta = -2.5$ and $\theta = 2.5$ increase. When $\delta = 0.5$ test lengths at the extremes typically differ by approximately two test items. When $\delta = 1.0$ test lengths at the extremes typically differ by approximately four to six items.

The most extreme example of the deterioration in symmetry occurs when three initial items are administered and $\text{s.e.}(\hat{\theta}) < 0.35$. When $\delta = 1.0$, 20.75 test items, on average, result for true $\theta = 0$. This increases to 24.21 and 31.51 test items, a 7.3 item difference, when θ moves out to the extreme values of -2.5 and 2.5, respectively. When $\text{s.e.}(\hat{\theta}) < 0.5$ and $\delta = 1$, 9.49 test items, on average, result for true $\theta = 0$. This increases to 13.16 and 19.53 test items, a 6.37 item difference, when true θ moves out to the extreme values of -2.5 and

2.5, respectively. Proportionally this represents a much larger difference in test length.

These disproportionate differences seem to be apparent throughout most testing conditions. For example, when $\delta = 0.5$ and $\text{s.e.}(\hat{\theta}) < 0.2$, differences in test length between $\theta = -2.5$ and $\theta = 2.5$ are approximately 4% to 5% of the average test length for average ability examinees, that is, where $\theta = 0$. When $\text{s.e.}(\hat{\theta}) < 0.35$ these differences are approximately 9% to 14% of the average test length for average ability examinees. These differences increase to approximately 14% to 25% of the average test length for examinees of average ability when $\text{s.e.}(\hat{\theta}) < 0.5$. When $\delta = 1.0$ differences are less than 10% of the average test length for average ability examinees when $\text{s.e.}(\hat{\theta}) < 0.2$. This jumps to approximately 25% to 35% when $\text{s.e.}(\hat{\theta}) < 0.35$ and dramatically increases to approximately 44% to 67% when $\text{s.e.}(\hat{\theta}) < 0.5$.

Another disproportional pattern that occurs in relation to the effect of the manipulation of item difficulty on test length has to do with the standard deviation of the number of test items, $\text{s.d.}(n)$. Tables 6, 7, and 8 present the average number of test items, averaging first over subjects then over θ , along with the corresponding average standard deviations. Also included is the percentage these standard deviations represent with respect to their corresponding average number of test items. These tables show that as δ increases, test length increases, as expected, and $\text{s.d.}(n)$ also increases. What is unexpected is that as δ increases, the increase in $\text{s.d.}(n)$ is not proportional to the increase in test length.

Table 6: Average number of test items (n), s.d.(n) and corresponding percent with s.e.(θ) of 0.2.

Delta	3 initial items			5 initial items			7 initial items		
	Average n	Average s.d.(n)	%	Average n	Average s.d.(n)	%	Average n	Average s.d.(n)	%
0	38.93	1.55	4.0%	39.83	1.52	3.8%	40.71	1.51	3.7%
0.1	39.15	1.85	4.7%	40.07	1.91	4.8%	40.88	1.57	3.8%
0.2	39.89	2.34	5.9%	40.83	2.34	5.7%	41.94	2.29	5.5%
0.3	41.05	2.84	6.9%	42.00	3.07	7.3%	42.80	2.72	6.4%
0.4	42.96	3.87	9.0%	43.75	3.92	9.0%	44.54	3.35	7.5%
0.5	45.49	4.90	10.8%	46.15	4.68	10.1%	46.98	4.62	9.8%
0.6	48.50	5.91	12.2%	49.11	5.78	11.8%	49.63	5.58	11.2%
0.8	56.93	8.69	15.3%	57.47	8.45	14.7%	57.59	7.84	13.6%
1.0	70.14	12.36	17.6%	69.80	12.23	17.5%	69.92	11.36	16.2%

Test items are averaged over subjects and ability levels.
Percentage represents: average s.d.(n) / average n

Table 7: Average number of test items (n), s.d.(n) and corresponding percent with s.e.(θ) of 0.35.

Delta	3 initial items			5 initial items			7 initial items		
	Average n	Average s.d.(n)	%	Average n	Average s.d.(n)	%	Average n	Average s.d.(n)	%
0	14.75	1.04	7.1%	15.59	1.15	7.4%	16.60	1.34	8.1%
0.1	14.86	1.19	8.0%	15.68	1.21	7.7%	16.63	1.36	8.2%
0.2	15.07	1.26	8.4%	15.89	1.33	8.4%	16.87	1.50	8.9%
0.3	15.46	1.51	9.8%	16.28	1.51	9.3%	17.18	1.63	9.5%
0.4	16.06	1.79	11.1%	16.77	1.77	10.6%	17.66	1.95	11.0%
0.5	16.73	2.14	12.8%	17.39	2.07	11.9%	18.19	2.16	11.9%
0.6	17.70	2.82	15.9%	18.33	2.58	14.1%	18.91	2.42	12.8%
0.8	20.50	4.37	21.3%	20.87	4.01	19.2%	21.15	3.62	17.1%
1.0	24.65	6.12	24.8%	24.59	5.51	22.4%	24.89	5.30	21.3%

Test items are averaged over subjects and ability levels.

Percentage represents: average s.d.(n) / average n

Table 8: Average number of test items (n), s.d.(n) and corresponding percent with s.e.(θ) of 0.5.

Delta	3 initial items			5 initial items			7 initial items		
	Average n	Average s.d.(n)	%	Average n	Average s.d.(n)	%	Average n	Average s.d.(n)	%
0	8.62	0.98	11.4%	9.54	1.21	12.7%	10.93	1.25	11.4%
0.1	8.63	0.96	11.1%	9.63	1.16	12.0%	10.88	1.29	11.9%
0.2	8.73	1.12	12.8%	9.68	1.25	12.9%	10.93	1.27	11.6%
0.3	8.89	1.06	11.9%	9.70	1.30	13.4%	10.94	1.33	12.2%
0.4	8.94	1.13	12.6%	9.80	1.36	13.9%	11.08	1.38	12.5%
0.5	9.32	1.26	13.5%	9.99	1.52	15.2%	11.26	1.48	13.1%
0.6	9.62	1.48	15.4%	10.37	1.70	16.4%	11.42	1.71	15.0%
0.8	10.78	2.09	19.4%	11.17	2.22	19.9%	12.22	2.16	17.7%
1.0	12.57	3.05	24.3%	12.72	2.99	23.5%	13.41	2.91	21.7%

Test items are averaged over subjects and ability levels.
Percentage represents: average s.d.(n) / average n

For $s.e.(\hat{\theta}) < 0.2$, regardless of the number of initial test items administered, the average s.d.(n) represents approximately 4% of the average number of test items administered to examinees when optimal items are administered, that is, when $\delta = 0$. As δ increases the percentage represented by the s.d.(n), as related to the average number of test items, steadily increases.

When $\delta = 1.0$, this percentage ranges from 16.2% to 17.6%, depending on the number of initial test items administered. When $s.e.(\hat{\theta}) < 0.35$, the average s.d.(n) ranges from 7.1% to 8.1%, for $\delta = 0$, depending on the number of initial test items administered. This increases to 21.3% to 24.8% for $\delta = 1.0$. When $s.e.(\hat{\theta}) < 0.5$, the average s.d.(n) ranges from 11.4% to 12.7%, for $\delta = 0$, depending on the number of initial test items administered. This increases to 21.7% to 24.3% for $\delta = 1.0$.

In general, as δ increases to the highest levels of modification, there is a great deal of variation in the resulting test lengths within each of the true θ values.

B. Number of initial test items administered

In this study three, five and then seven initial test items (n), respectively, were administered prior to the start of the adaptive portion of the test, for each major version of the study. Tables 9, 10, and 11 present the average number of test items required to achieve each of the desired levels of precision considered in this study. For almost every δ value, regardless of the number of initial test

Table 9: Average number of test items (n) with $s.e.(\hat{\theta}) < 0.2$

δ	<u>3 initial items</u>	<u>5 initial items</u>	<u>7 initial items</u>
0	38.93	39.83	40.71
0.1	39.15	40.07	40.88
0.2	39.89	40.83	41.76
0.3	41.05	42.00	42.80
0.4	42.96	43.75	44.54
0.5	45.49	46.15	46.98
0.6	48.50	49.11	49.63
0.8	56.93	57.47	57.59
1.0	70.14	69.80	69.92

Table 10: Average number of test items (n) with $s.e.(\hat{\theta}) < 0.35$

δ	<u>3 initial items</u>	<u>5 initial items</u>	<u>7 initial items</u>
0	14.75	15.59	16.60
0.1	14.86	15.68	16.63
0.2	15.07	15.89	16.87
0.3	15.46	16.28	17.18
0.4	16.06	16.77	17.66
0.5	16.73	17.39	18.19
0.6	17.70	18.33	18.91
0.8	20.50	20.87	21.15
1.0	24.65	24.59	24.89

Table 11: Average number of test items (n) with $s.e.(\hat{\theta}) < 0.5$

<u>δ</u>	<u>3 initial items</u>	<u>5 initial items</u>	<u>7 initial items</u>
0	8.62	9.54	10.93
0.1	8.63	9.63	10.88
0.2	8.73	9.68	10.93
0.3	8.89	9.70	10.94
0.4	8.94	9.80	11.08
0.5	9.32	9.99	11.26
0.6	9.62	10.37	11.42
0.8	10.78	11.17	12.22
1.0	12.57	12.72	13.41

items administered, there is less than a one-item difference, on average, except for $s.e.(\hat{\theta}) < 0.5$ when comparing the conditions of five and seven initial test items.

For example, as presented in Table 9, when $\delta = 0$ and $s.e.(\hat{\theta}) = 0.2$ the average number of test items is 38.93 when three initial test items are administered. This increases to 39.83 test items when five initial test items are administered. When seven initial test items are administered the average number of test items increases to 40.71.

When $s.e.(\hat{\theta}) < 0.5$ the average number of test items increases by more than one item, over all but one of the δ values, when comparing the condition where five initial test items are administered to the condition where seven initial items are administered.

As δ reaches its highest values, in several cases there is less than a one-item difference, on average, between the three-initial item version and the seven-initial item version. For example, this occurs when $\text{s.e.}(\hat{\theta}) < 0.2$ and 0.35 and $\delta = 0.8$ and 1.0 . Here, as presented in Table 11, the average number of test items is 12.57 when three initial test items are administered. This increases to 12.72 test items when five initial test items are administered. When seven initial test items are administered, the average number of test items increases to 13.41 .

When examining a graphical representation of this data in Figures 15, 16 and 17, one can easily observe the close proximity of the data points under the different levels of precision. It is also easy to see the narrowing of the distance between these lines to a point where the lines are almost coincident for $\text{s.e.}(\theta)$ of 0.2 , 0.35 , as well as 0.5 .

Figure 15: Average number of test items (n) with s.e.(theta) of 0.2

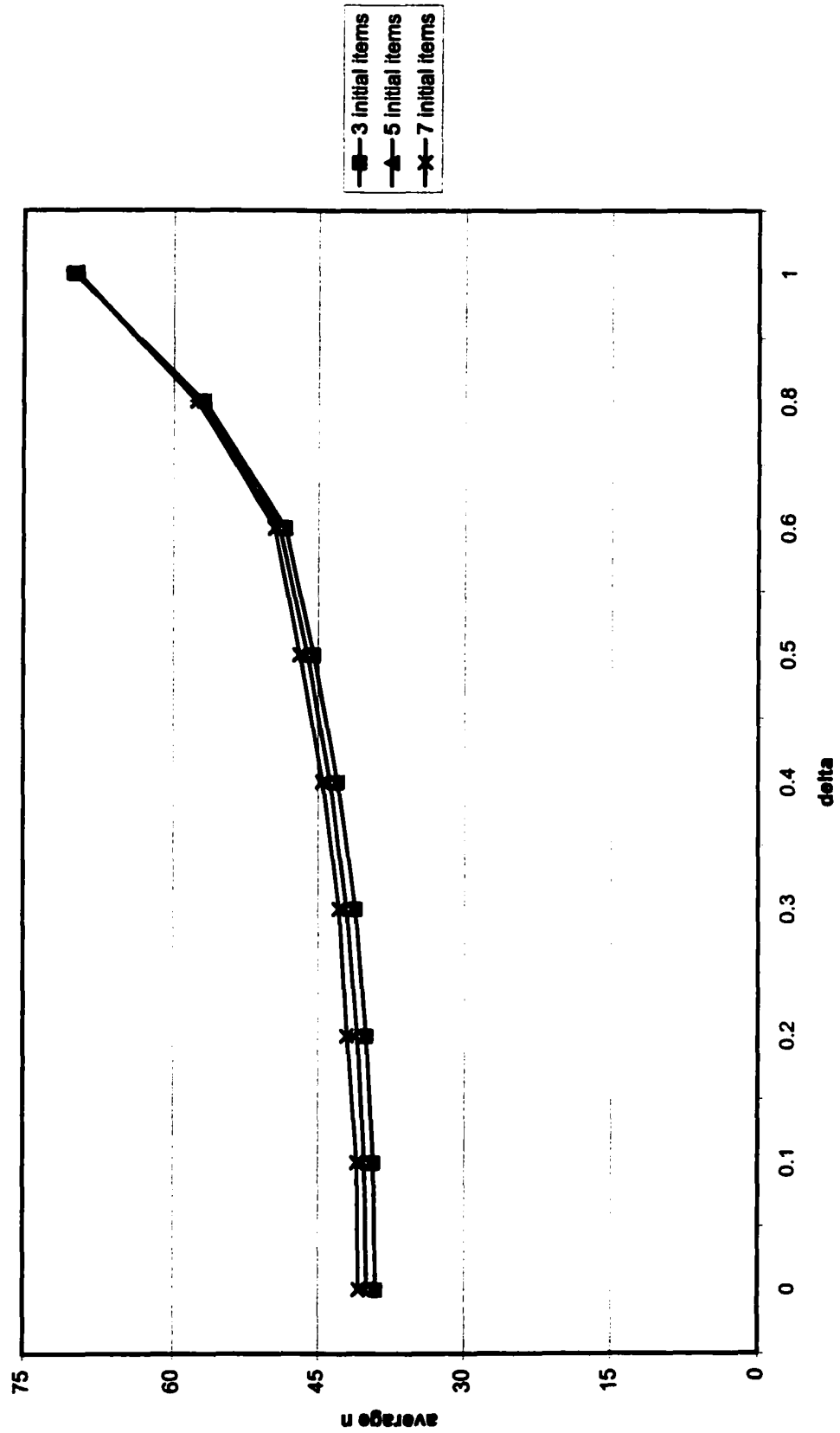


Figure 16: Average number of test items (n) with $s.e.(\theta)$ of 0.35

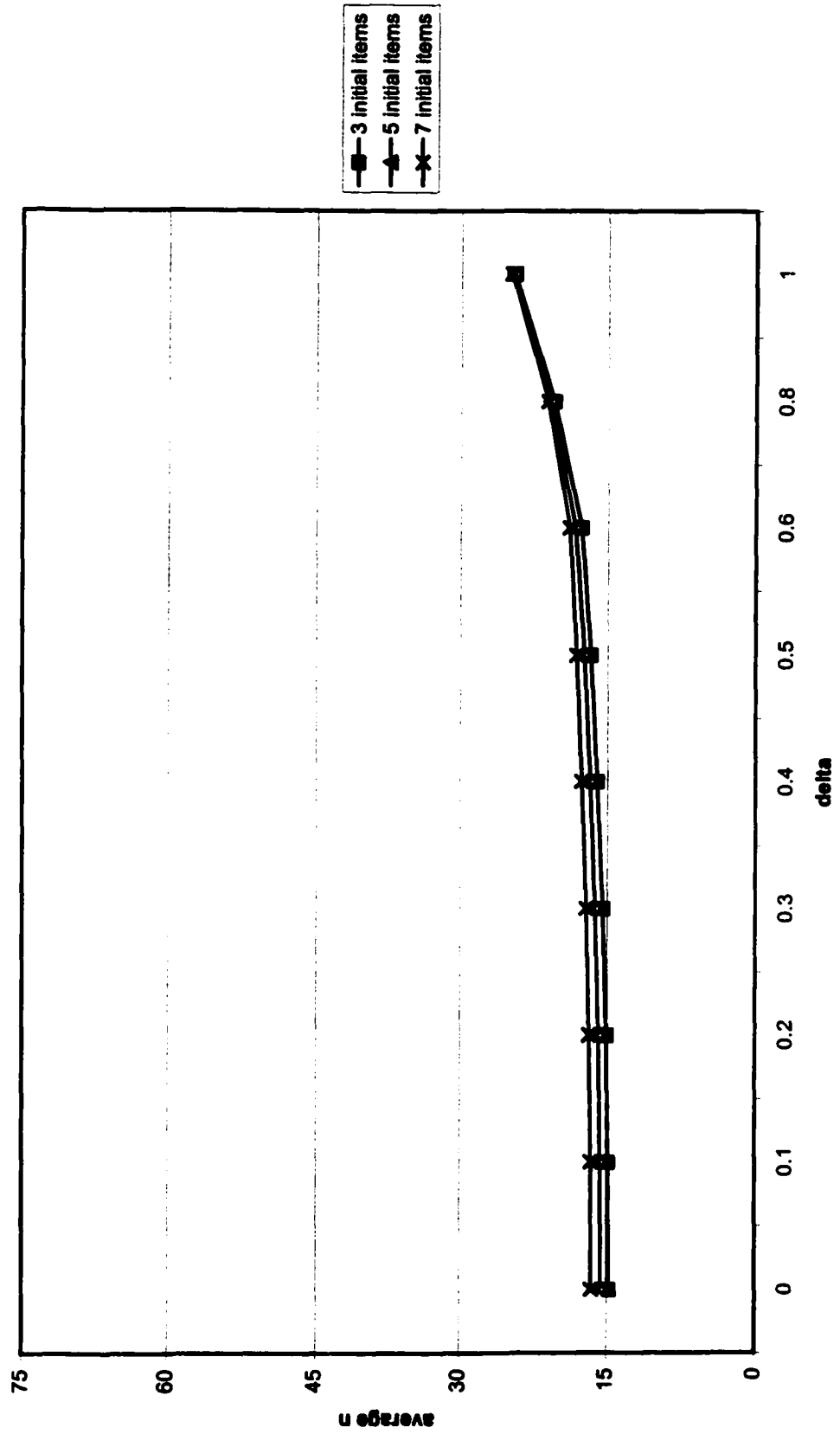
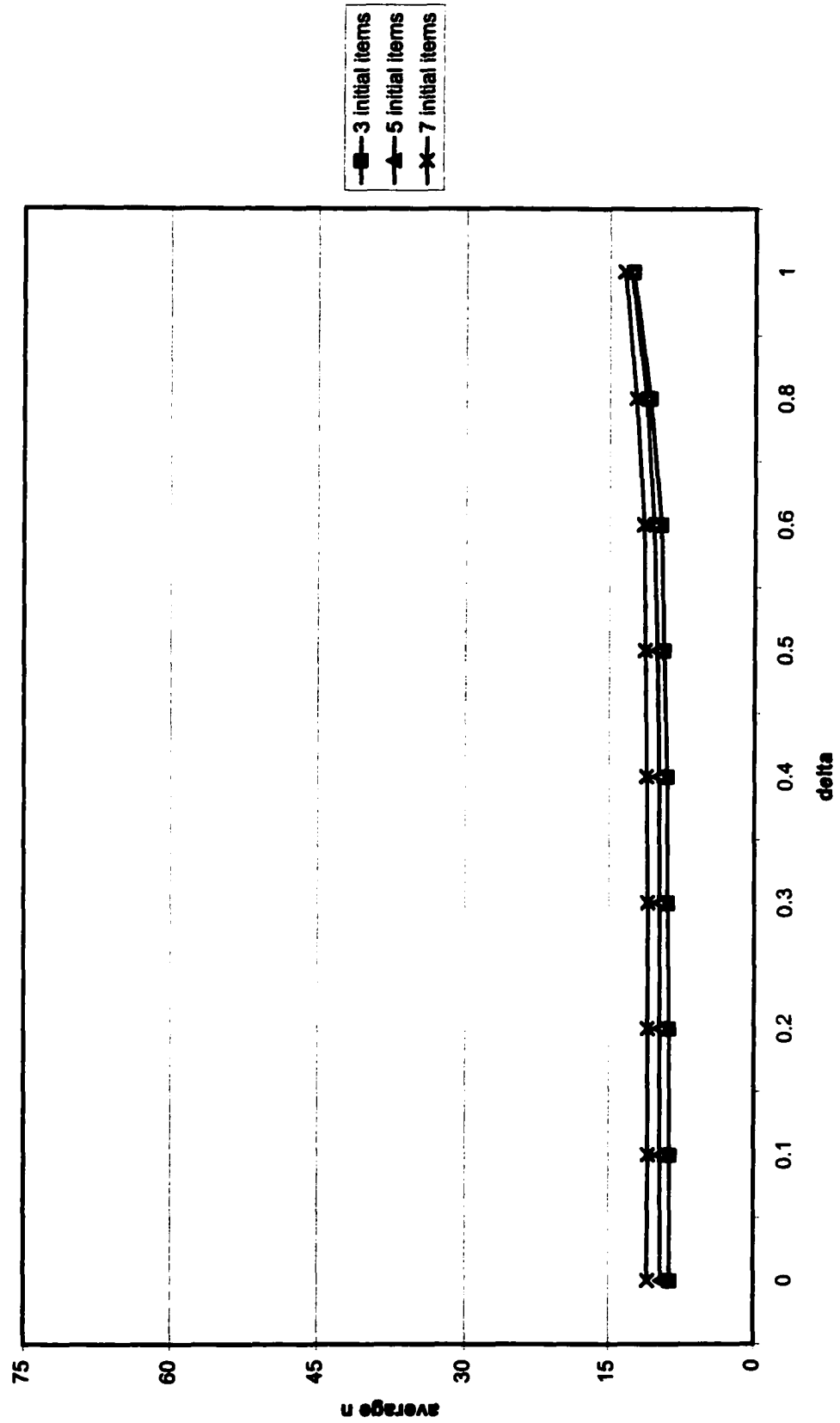


Figure 17: Average number of test items (n) with s.e.(theta) of 0.5



C. Ability estimates

Upon examination of the data presented in Appendix C one can see an illustration of the property of parameter invariance. This property indicates that θ is expected to be estimated with a relatively high degree of accuracy regardless of the modifications made to the levels of item difficulty in this study. In all of the circumstances considered in this study this seems to be the case. All of the ability estimates very closely approximate the corresponding true θ value. In fact, 113 of the 891 ability estimates differ from their corresponding true θ values by less than 0.005.

As $s.e.(\theta)$ increases, thus permitting more measurement error, there is some decrease in the accuracy of these θ estimates. Additionally, as one examines the data presented in Appendix C, one can see that the standard deviations of the ability estimates, $s.d.(\theta)$, all are relatively close to the theoretical $s.e.(\theta)$ values which were preset as the stopping rules for the different versions of the computer simulation.

Chapter V

SUMMARY AND DISCUSSION

For this study a series of computer simulations were conducted to determine the effect of the manipulation of item difficulty on test length in a CAT environment. Also under examination was the effect of administering different numbers of initial test items, prior to the start of adaptive portion of the test, on test length. Finally, the effect, if any, on the consistency of ability estimates with respect to the two previous modifications has been considered.

A. Ability Estimates

With all the modifications of traditional CATs made in this study, the procedures seem to result in unbiased ability estimates. This and the fact that the standard deviations of the ability estimates, $s.d.(\hat{\theta})$, closely approximate the theoretical $s.e.(\theta)$ values studied here, shows that the procedures undertaken here work well.

B. Number of Initial Test Items Administered

There is very little impact on test length when comparing the use of three, five or seven initial items prior to the start of the adaptive portion of a CAT. Unless the examinees are all of average ability there seems to be no advantage

to adding extra items of average difficulty to the beginning of the exam. These items would not provide a great deal of information about examinees whose abilities lie further from the middle of the ability continuum.

It would seem prudent to use fewer initial test items, rather than more, prior to the start of the adaptive portion of the test. This would enable test makers to preserve the integrity of the test by exposing examinees to fewer average-difficulty items, in general, because all examinees are exposed to this level of difficulty upon starting the exam. It would seem unnecessary, and some may even say unreasonable, to use five or seven initial items on a CAT. This may place an unnecessary burden on test makers as well as examinees. If it were known that all examinees were to receive five or seven initial items all of average difficulty, test makers would have to increase the item pool sufficiently to avoid overexposure of these items. Considering this, the use of three initial items seems to be the most reasonable choice.

C. Manipulation of Item Difficulty

The most important effect under consideration here is the impact of the use of non-optimal items on a CAT. Under the most precise condition where $\text{s.e.}(\hat{\theta}) < 0.2$ and the optimal psychometric condition of $\delta = 0$, average test length is 38.93 to 40.71 items, depending on the number of initial test items administered. When considering the modifications made to item difficulty there is almost no noticeable impact on test length when using $\delta = 0.1$ or 0.2 . These

modifications only result in the modest increase in the average test length of approximately one item, at most, which represents approximately a 2.5% increase in test length. When $\delta = 0.3$ there is still only a modest increase in test length of two test items, about a 5% increase.

Some test makers or administrators may even consider using $\delta = 0.4$. Here average test length increases by 3.83 to 4.03 test items (a 9.4% to 10% increase depending on the number of initial test items administered). This may seem like a reasonable concession when considering the potential psychological advantages of having the opportunity to answer easier test items than those offered on a traditional CAT. In this situation, an examinee's chances of a correct response increase from 50%, where optimal CAT items are employed, to 66.4%.

When $\delta = 0.5$ the probability of an examinee correctly responding to a test item increases to 70.1%. Here average test length increases by 6.27 to 6.56 test items, depending on the number of initial test items administered. This represents an increase of 15.4% to 16.9%. This increase in test length, though large, may be justifiable in some situations when considering the potential psychological advantages.

When $\delta = 0.6$ the probability of an examinee correctly responding to a test item increases to 73.5%. Here average test length increases by 8.92 to 9.57 test items, depending on the number of initial test items administered. This represents an increase of 21.9% to 24.6%. This may seem like a large increase in test length, but in some situations, may be warranted. Recall that CATs can

be as little as half as long as tests based on classical test theory, so even when $\delta = 0.6$ there still may be a considerable savings in test length on the CAT. With certain examinees test anxiety has the potential to negatively affect their test scores. CATs using psychometrically optimal items ($\delta = 0$) have the potential to increase this test anxiety, given that an individual only has a 50% chance of a correct response. In a situation where an examinee has severe test anxiety, administering items where $\delta = 0.6$ may be justified in order to give the examinee the best possible testing situation by lessening the impact of this potentially stressful testing situation.

As δ increases to 0.8, test length increases by approximately 17 to 18 items. This represents an approximate increase of 42% to 46%, depending on the number of initial test items. Few would expect such an increase to be considered reasonable. Certainly $\delta = 1.0$ creates an unacceptable situation. Here test length increases by 29.21 to 31.21 test items, depending on the number of initial test items. This represents a 71.8% to 80.2% increase in test length.

When $\text{s.e.}(\hat{\theta}) < 0.35$ and $\delta = 0$, average test length is 14.75 to 16.60 items, depending on the number of initial test items. When considering the modifications made to item difficulty, δ values of 0.1, 0.2, and 0.3 all result in less than a one-item increase in test length. When $\delta = 0.4$ there is just over a one item increase. Based on these results none of these modifications seem unreasonable.

When $\delta = 0.5$ there is only a 1.59 to 1.98 item increase in test length, on average. This represents a 9.6% to 13.4% increases in test length. When $\delta = 0.6$ test length only increases, on average, by 2.31 to 2.95 test items, a 13.9% to 20% increase. Even with these increases in test length, these modifications seem reasonable given the potential psychological advantages.

Some test makers or administrators may even consider increasing average test length by 4.55 to 5.77 test items (a 27.4% to 39.1% increase), when $\delta = 0.8$, a reasonable concession when considering the potential psychological advantages. Recall that here an examinee's chances of a correct response increase from 50%, in the optimal CAT condition, to 79.6%.

When $\delta = 1.0$ the probability of an examinee correctly responding to a test item increases to 84.6%. Here average test length increases by 8.29 to 9.89 test items, depending on the number of initial test items administered. This represents an increase of 49.9% to 67.1%. This may seem like an extremely large increase in test length, especially considering that test length under optimal CAT conditions are approximately 15 to 17 items, and may not be considered justifiable.

Under the least precise condition under consideration here, $\text{s.e.}(\hat{\theta}) < 0.5$, when $\delta = 0$ average test length is only 8.62 to 10.93 items, depending on the number of initial items administered. Recall that this level of precision is likely to occur with typical teacher-made tests and probably would not be the case for a large-scale testing situation. When considering all possible modifications to item difficulty, regardless of the number of initial test items administered, all resulting

increases to test length are at most one item for modifications up to $\delta = 0.6$.

Here increases in test length represent at most 11.6%. Considering these results in absolute terms, any one of these modifications seems completely reasonable if one were willing to accept this level of precision for ability estimation.

When $\delta = 0.8$, test length only increases by 1.29 to 2.16 test items (a 2.5% to 11.8% increase). Recall that the probability of a correct response increases to 79.6% in this situation. Even when $\delta = 1.0$, and the probability of a correct response increases to 84.6%, test length only increases by 2.48 to 3.95 items, on average. This represents a 22.7% to 45.8% increase. Given the potential psychological advantages here, some may find it very easy to argue for the examinees and use such easy items. When tests are relatively short to begin with, examinees probably would be more than willing to take a slightly longer test made up of easier items than a shorter test of optimal, and therefore, more difficult items. Again, these modifications seem reasonable to accept providing one is willing to accept the corresponding level of precision.

D. Implications

In this study it has been shown that modifications to test item difficulty may be made without having a major impact on the estimation of ability. In addition, there are conditions where some of these modifications have little practical impact on test length. When considering the potential negative psychological impact of using optimal CAT items, some test administrators may consider

modifying test item difficulty as discussed throughout this study. For certain examinees extreme test anxiety may be a factor which may impact not only on the testing situation, but possibly on test scores as well. There are modifications that can be made which may result in relatively large increases in test length. Some test administrators may be willing to make these concessions considering the potential benefits to such an examinee. Chances are likely that such exams would still have fewer items than exams written using classical testing methodology.

It is important to realize that the results reported for modifications at each δ value are averages over all examinees at each particular true θ value. At the higher δ values there was a disproportionately larger s.d.(n) than at the lower δ values. This means that individual scores, within any particular true θ value, will vary greatly. So this must also be taken in consideration when applying any of the modifications to item difficulty presented in this study.

Another consideration when modifying item difficulty is the effect on item pools. Not only will test designers need to consider the potential increases in test length, possibly necessitating a larger item pool, but also the difficulties of the items required for such modified tests. As modifications to item difficulty are implemented, test items become easier based on those modifications, causing a shift in the distribution of item difficulties required. That is, when designing a CAT using optimal items, a majority of the items would probably be of average difficulty, assuming a normally distributed population of examinees' abilities. If

items are made easier, the majority of the items would now be easier than average, by a factor of δ , that is, based on the modifications implemented.

E. Future Research

Future investigations are recommended to assess the potential psychological impact on examinees when administering optimal CAT items. Studies should also assess the impact of the modifications suggested in this study relating to test item difficulty on: stress levels during the CAT, post-test anxiety levels, ability estimation and test length. Examinees could be randomly assigned to different test conditions, each implementing different levels of δ . Assuming the ability distributions are the same in all of the groups, based on random assignment, the ability estimates should be the same in all of the groups. If not, this would provide some initial evidence that the more difficult test items may be negatively affecting ability estimates, assuming the more difficult items result in significantly lower ability estimates. Higher stress levels reported by examinees taking an optimal CAT, as compared to CATs implementing easier items, would provide evidence for administering easier CAT items.

Different levels of δ may be studied with examinees exhibiting various levels of test anxiety, where this anxiety has the potential to have an almost paralyzing effect in the more extreme cases. Again, examinees could be randomly assigned to different test conditions, each implementing different levels of δ . The ability estimates should be the same in all of the groups. If not, this would again provide some evidence that the more difficult test items may be

negatively affecting ability estimates, assuming the more difficult items result in significantly lower ability estimates. In addition, anxiety inventories could be administered to examinees in all δ conditions to determine the effect of manipulating item difficulty on test anxiety.

Additional studies could be conducted to determine the impact of allowing examinees to choose the level of difficulty for test items on anxiety levels, ability estimation, and test length in a CAT environment. This would allow for multiple levels of modifications to item difficulty to be made throughout individual exams. This could be implemented with examinees choosing levels of item difficulty based on a Likert scale corresponding to different levels of δ within an adaptive framework.

In addition to the research suggested above, additional studies could be conducted to address the impact of employing different levels of choice. For example, examinees could be offered the choice of 2, 3 or possibly more levels of modifications to item difficulty, depending on the study conditions to which they are assigned.

In general, more research concerning the use of optimal and non-optimal items in a CAT environment is needed. As CAT is implemented more and more for mass test administration, the positive as well as the negative aspects will come to light. The modifications suggested in this study provide a starting point for addressing one of the potential negative aspects.

APPENDIX A

Qbasic program to simulate test results for CAT exams

```

REM CAT simulation
REM ***Now startup sets up output file
CLS
filename$ = "A:output.txt"
CALL Choose(filename$)
OPEN filename$ FOR OUTPUT AS #2
PRINT #2, " "
PRINT #2, " "
REM
REM
DIM B(1 TO 1000), P(1 TO 1000)
trutheta = -3
REM ***** note change for each run *****
delta = 0
PRINT #2, "delta = "; delta
REM x = # of subjects in each group
x = 100
FOR g = 1 TO 11
  trutheta = trutheta + .5
  sumitems = 0
  sumtheta = 0
  sumthsq = 0
  sumitmsq = 0
  PRINT #2, " "
  PRINT #2, "True theta="; trutheta

  FOR j = 1 TO x
    R = 0
    REM administer 1st 3 items at B = 0 to get initial estimate of theta
    n = 3
    FOR i = 1 TO n
      B(i) = 0
      u = RND
      REM calculate subject's actual chance of answering this item correctly
      prob = 1 / (1 + EXP(-1.7 * (trutheta - B(i))))
      IF u < prob THEN
        answer = 1
      ELSE
        answer = 0
      END IF
      R = R + answer
    NEXT i
  
```

```

REM calculate initial closed form estimate of theta from R and i
IF R = 0 OR R = n THEN
  theta = (LOG((R + .5) / (n - R + .5))) / 1.7
ELSE
  theta = (LOG(R / (n - R))) / 1.7
END IF

```

```

REM continuing with item 4 to end of test for subject 1...

```

```

DO
  n = n + 1
  B(n) = theta - delta
  u = RND
  prob = 1 / (1 + EXP(-1.7 * (trutheta - B(n))))
  IF u < prob THEN
    answer = 1
  ELSE
    answer = 0
  END IF
  R = R + answer

```

```

IF R = 0 THEN
  R = R + .5
END IF
IF R = n THEN
  R = R - .5
END IF

```

```

REM enter into N - R iterative procedure with appropriate R

```

```

DO
  SP = 0
  spq = 0
  FOR i = 1 TO n
    P(i) = EXP(1.7 * (theta - B(i))) / (1 + EXP(1.7 * (theta - B(i))))
    SP = SP + P(i)
    spq = P(i) * (1 - P(i)) + spq
  NEXT i
  H = (R - SP) / (-1.7 * spq)
  theta = theta - H
  LOOP UNTIL ABS(H) < .001

```

```

seth = 1 / SQR(1.7 ^ 2 * spq)

```

REM reset R back to 0 or n as necessary

IF R = .5 THEN

R = 0

END IF

IF R = n - .5 THEN

R = R + .5

END IF

LOOP UNTIL seth < .2 OR n = 200

sumtheta = sumtheta + theta

sumitems = sumitems + n

sumthsq = sumthsq + (theta ^ 2)

sumitmsq = sumitmsq + (n ^ 2)

NEXT j

PRINT #2, "Average theta = "; sumtheta / x

PRINT #2, "Average # items = "; sumitems / x

PRINT #2, "std dev of theta = "; SQR((sumthsq - (sumtheta ^ 2) / x) / (x - 1))

PRINT #2, "sd # of items = "; SQR((sumitmsq - ((sumitems ^ 2) / x)) / (x - 1))

NEXT g

REM

REM NOW return to startup routine for output file

CLOSE #2

END

SUB Choose (filename\$)

REM ***allows user to choose output to screen or disk file

,

INPUT " Output to screen or disk (S or D)", ans\$

IF UCASE\$(LEFT\$(ans\$, 1)) = "S" THEN

filename\$ = "con"

END IF

CLS

END SUB

APPENDIX BAverage number of test items (n) for true theta (θ) and corresponding s.d. with 3 initial test items s.e. (θ) of 0.2

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	40.86 (1.69)	41.18 (1.95)	41.33 (1.77)	42.45 (3.09)	44.20 (3.98)	46.98 (5.05)	49.00 (5.75)	58.14 (8.19)	69.53 (11.26)
-2.0	40.01 (1.64)	39.9 (1.59)	40.37 (2.04)	41.87 (2.43)	43.40 (3.60)	46.31 (4.43)	48.34 (6.06)	56.24 (8.30)	69.20 (13.28)
-1.5	39.37 (1.76)	39.31 (2.43)	39.98 (2.35)	41.61 (3.19)	43.80 (4.34)	45.85 (5.24)	49.36 (5.96)	56.38 (9.78)	72.44 (13.15)
-1.0	38.36 (1.71)	38.48 (1.50)	39.40 (2.61)	40.78 (3.24)	41.80 (4.13)	44.38 (4.62)	47.88 (6.12)	56.42 (9.19)	70.18 (13.49)
-0.5	37.29 (1.42)	37.37 (1.59)	38.18 (2.23)	39.00 (2.45)	41.11 (3.19)	43.32 (4.25)	45.88 (5.34)	56.79 (8.28)	66.25 (14.28)
0	37.13 (1.57)	37.05 (2.01)	37.65 (1.74)	38.87 (3.40)	40.27 (3.75)	43.13 (4.92)	46.10 (5.29)	54.40 (8.87)	66.02 (12.42)
0.5	37.06 (1.17)	37.37 (1.80)	38.51 (2.72)	39.03 (2.48)	41.46 (4.09)	43.27 (4.85)	45.79 (5.73)	52.33 (7.24)	67.35 (12.54)
1.0	38.38 (1.76)	38.44 (1.76)	39.34 (2.44)	40.00 (2.13)	43.03 (4.47)	45.64 (5.16)	47.23 (6.20)	54.87 (8.89)	68.84 (11.73)
1.5	39.13 (1.42)	39.52 (1.76)	40.56 (2.72)	41.58 (2.94)	43.43 (4.35)	45.6 (4.92)	49.59 (5.97)	58.46 (8.53)	72.05 (11.32)
2.0	40.01 (1.81)	40.49 (2.19)	41.23 (3.11)	42.76 (2.74)	44.04 (3.05)	46.85 (5.59)	50.86 (5.92)	61.15 (10.25)	73.94 (11.31)
2.5	40.63 (1.08)	41.56 (1.79)	42.25 (1.99)	43.59 (3.19)	46.07 (3.59)	49.11 (4.90)	53.45 (6.67)	61.07 (8.09)	75.79 (11.15)

Entries include average # of test items (n) and corresponding standard deviations listed below in parentheses.

Average number of test items (n) for true theta (θ) and corresponding s.d. with 5 initial test items s.e. (θ) of 0.2

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	42.42 (1.51)	42.72 (2.18)	43.35 (2.52)	44.54 (3.27)	45.47 (3.29)	47.18 (4.35)	51.50 (6.20)	59.98 (9.27)	72.10 (13.82)
-2.0	41.48 (1.55)	41.55 (1.65)	42.41 (2.18)	43.85 (3.18)	45.10 (3.16)	47.32 (4.41)	50.36 (6.47)	59.35 (8.43)	72.00 (12.48)
-1.5	40.86 (1.98)	40.73 (2.09)	41.73 (2.68)	42.96 (3.10)	44.45 (4.73)	47.06 (4.41)	51.08 (6.82)	58.21 (9.39)	73.45 (13.40)
-1.0	39.05 (1.51)	39.15 (1.62)	40.07 (2.85)	41.58 (3.23)	43.77 (5.16)	45.82 (5.50)	48.93 (5.25)	56.93 (8.26)	68.61 (13.47)
-0.5	37.39 (1.49)	37.53 (1.70)	38.54 (2.41)	39.11 (2.86)	40.70 (3.74)	43.80 (4.73)	46.48 (5.15)	55.21 (8.56)	66.50 (12.05)
0	36.38 (0.96)	36.49 (1.17)	37.37 (2.05)	38.48 (2.75)	40.24 (3.20)	42.54 (4.38)	44.71 (5.31)	52.26 (7.66)	63.57 (11.98)
0.5	37.67 (1.64)	37.63 (1.71)	38.15 (2.26)	39.06 (3.01)	40.95 (3.55)	43.29 (4.50)	45.04 (5.27)	53.31 (7.70)	65.96 (12.99)
1.0	38.71 (1.56)	39.10 (2.19)	39.75 (2.32)	41.07 (3.01)	42.16 (3.03)	46.14 (5.05)	47.12 (6.04)	55.04 (7.92)	67.54 (10.82)
1.5	40.28 (1.53)	40.79 (1.64)	41.59 (2.46)	42.53 (3.47)	44.94 (5.08)	46.80 (4.94)	49.89 (5.57)	59.00 (8.63)	68.92 (10.38)
2.0	41.41 (1.49)	42.07 (2.52)	42.29 (1.83)	44.27 (3.62)	46.15 (3.84)	48.32 (4.89)	51.36 (5.86)	60.92 (8.74)	72.79 (11.28)
2.5	42.50 (1.51)	42.94 (2.30)	43.91 (2.22)	44.60 (2.31)	47.34 (4.37)	49.36 (4.32)	53.73 (5.67)	61.98 (8.39)	76.35 (11.91)

Entries include average # of test items (n) and corresponding standard deviations listed below in parentheses.

Average number of test items (n) for true theta (θ) and corresponding s.d. with 7 initial test items s.e. (θ) of 0.2

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	43.85 (1.34)	44.29 (2.08)	45.14 (2.70)	45.68 (2.65)	47.47 (4.29)	50.08 (5.55)	53.62 (8.08)	62.44 (9.46)	76.04 (14.95)
-2.0	43.06 (1.72)	43.09 (1.74)	43.95 (2.09)	45.03 (2.65)	46.85 (3.26)	49.20 (4.31)	52.35 (5.31)	59.92 (8.30)	73.35 (11.00)
-1.5	41.86 (1.44)	41.88 (1.52)	43.07 (2.64)	44.11 (3.23)	46.19 (3.72)	49.32 (5.58)	51.98 (6.14)	59.74 (8.96)	73.57 (12.16)
-1.0	39.49 (1.45)	39.94 (1.57)	40.97 (2.87)	41.89 (2.57)	43.86 (3.97)	45.62 (4.66)	47.95 (5.22)	57.14 (8.15)	69.19 (11.85)
-0.5	37.29 (1.21)	37.51 (1.53)	38.41 (2.28)	39.55 (2.88)	40.77 (3.15)	43.86 (5.01)	45.84 (5.05)	53.48 (7.67)	65.37 (11.95)
0	36.03 (0.81)	36.35 (1.07)	36.97 (1.57)	38.28 (2.47)	39.51 (2.71)	42.01 (4.02)	43.93 (4.07)	51.51 (7.33)	62.04 (9.66)
0.5	37.74 (2.06)	37.57 (1.42)	37.97 (1.48)	39.01 (2.14)	40.48 (3.06)	42.77 (3.69)	46.15 (5.22)	51.53 (5.77)	62.76 (9.92)
1.0	39.66 (1.49)	39.70 (1.20)	40.71 (2.30)	41.54 (2.34)	43.20 (3.35)	44.98 (4.53)	47.38 (5.47)	54.61 (7.00)	67.07 (11.09)
1.5	41.78 (1.68)	41.88 (1.55)	42.96 (2.33)	43.97 (3.36)	45.35 (3.69)	47.61 (4.17)	51.63 (5.81)	58.83 (8.09)	70.14 (10.51)
2.0	43.06 (1.78)	43.25 (1.72)	44.36 (2.72)	45.05 (2.69)	47.41 (3.61)	49.69 (4.68)	51.75 (5.28)	62.03 (8.09)	74.16 (10.91)
2.5	43.98 (1.63)	44.27 (1.89)	44.88 (2.22)	46.71 (2.89)	48.88 (1.99)	51.65 (4.62)	55.30 (5.77)	62.28 (7.37)	75.45 (10.92)

Entries include average # of test items (n) and corresponding standard deviations listed below in parentheses.

Average number of test items (n) for true theta (θ) and corresponding s.d. with 3 initial test items s.e. (θ) of 0.35

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	16.73 (1.24)	16.54 (1.19)	16.45 (1.09)	16.85 (1.41)	17.45 (1.59)	18.30 (2.50)	18.94 (2.97)	20.94 (4.03)	24.21 (6.11)
-2.0	16.09 (1.26)	15.76 (1.30)	16.10 (1.28)	16.26 (1.75)	16.84 (1.99)	17.58 (2.28)	17.87 (2.25)	20.11 (3.93)	24.50 (6.13)
-1.5	15.07 (1.11)	15.05 (1.20)	15.26 (1.16)	15.61 (1.46)	16.27 (1.99)	16.61 (2.09)	17.40 (2.85)	20.64 (4.43)	24.50 (6.50)
-1.0	14.03 (0.96)	14.12 (1.09)	14.04 (1.06)	14.62 (1.36)	15.35 (1.67)	16.08 (2.27)	16.84 (2.66)	19.07 (4.11)	23.42 (5.75)
-0.5	13.05 (0.94)	13.15 (1.18)	13.41 (1.21)	13.61 (1.34)	14.23 (1.67)	14.96 (1.91)	16.02 (2.48)	18.35 (4.21)	21.56 (5.87)
0	12.76 (0.70)	12.79 (0.82)	13.00 (1.12)	13.35 (1.38)	13.90 (1.71)	14.38 (1.81)	15.09 (2.10)	17.73 (4.29)	20.75 (5.13)
0.5	13.10 (0.95)	13.30 (1.15)	13.29 (1.22)	13.75 (1.38)	14.05 (1.31)	14.79 (1.95)	15.88 (2.98)	17.57 (3.36)	21.88 (5.65)
1.0	14.08 (1.07)	14.47 (1.23)	14.38 (1.18)	15.02 (1.80)	15.35 (1.80)	15.50 (1.39)	16.35 (2.39)	19.75 (4.33)	24.21 (7.22)
1.5	14.93 (0.99)	15.10 (1.15)	15.34 (1.30)	15.65 (1.45)	16.15 (1.67)	16.80 (2.11)	18.88 (3.84)	21.57 (5.02)	25.62 (6.12)
2.0	15.58 (0.99)	16.04 (1.42)	16.76 (1.80)	16.84 (1.48)	17.88 (2.10)	18.68 (2.30)	19.69 (3.37)	23.59 (5.26)	28.95 (6.91)
2.5	16.86 (1.21)	17.15 (1.38)	17.75 (1.43)	18.48 (1.78)	19.24 (2.16)	20.31 (2.95)	21.73 (3.08)	26.38 (5.15)	31.51 (5.98)

Entries include average # of test items (n) and corresponding standard deviations listed below in parentheses.

Average number of test items (n) for true theta (θ) and corresponding s.d. with 5 initial test items s.e. (θ) of 0.35

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	18.24 (1.26)	18.25 (1.13)	18.06 (1.15)	18.73 (1.38)	19.04 (1.61)	19.59 (2.06)	20.38 (2.26)	22.86 (3.89)	26.85 (6.19)
-2.0	17.10 (1.09)	17.18 (1.08)	17.35 (1.37)	17.92 (1.41)	18.21 (1.78)	18.96 (2.21)	19.83 (2.51)	22.18 (3.44)	26.24 (5.01)
-1.5	16.00 (1.20)	16.14 (1.07)	16.27 (1.41)	16.68 (1.58)	17.22 (1.63)	18.39 (2.39)	18.97 (3.13)	21.34 (4.29)	25.36 (6.39)
-1.0	14.82 (1.46)	14.97 (1.66)	15.31 (1.57)	15.44 (1.63)	15.88 (1.93)	16.27 (2.00)	17.54 (2.64)	19.79 (4.65)	23.38 (4.93)
-0.5	13.01 (0.97)	13.12 (1.07)	13.37 (1.34)	13.66 (1.45)	14.25 (1.79)	14.99 (2.25)	15.52 (2.30)	18.16 (4.29)	20.70 (4.98)
0	12.51 (0.85)	12.49 (0.82)	12.65 (0.90)	12.96 (1.17)	13.48 (1.52)	13.82 (1.34)	14.33 (1.65)	16.13 (2.41)	18.10 (3.14)
0.5	13.40 (1.16)	13.35 (1.10)	13.41 (1.19)	13.65 (1.31)	13.96 (1.24)	14.66 (1.51)	15.17 (1.59)	16.57 (2.61)	20.02 (4.38)
1.0	15.05 (1.44)	14.94 (1.46)	15.21 (1.37)	15.33 (1.46)	15.60 (1.43)	15.90 (1.83)	16.98 (2.62)	20.09 (4.29)	23.02 (6.31)
1.5	16.10 (0.96)	16.16 (1.06)	16.49 (1.41)	16.73 (1.67)	17.20 (2.03)	17.67 (2.62)	19.27 (3.37)	21.54 (4.52)	25.91 (6.44)
2.0	17.07 (1.05)	17.25 (1.33)	17.69 (1.50)	18.24 (1.68)	19.01 (2.12)	19.61 (2.09)	21.23 (3.33)	24.44 (5.41)	29.08 (6.89)
2.5	18.23 (1.20)	18.66 (1.51)	19.00 (1.44)	19.76 (1.83)	20.63 (2.42)	21.40 (2.46)	22.36 (3.02)	26.42 (5.28)	31.71 (5.90)

Entries include average # of test items (n) and corresponding standard deviations listed below in parentheses.

Average number of test items (n) for true theta (θ) and corresponding s.d. with 7 initial test items s.e.(θ) of 0.35

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	19.84 (1.02)	19.70 (1.12)	20.03 (1.26)	20.22 (1.37)	20.59 (1.66)	21.41 (2.44)	22.20 (2.62)	24.91 (4.24)	29.13 (7.33)
-2.0	18.77 (1.34)	18.76 (1.22)	18.85 (1.35)	19.35 (1.41)	20.24 (3.26)	20.31 (1.89)	20.85 (2.28)	24.47 (3.63)	28.04 (6.66)
-1.5	17.33 (1.42)	17.42 (1.38)	17.83 (1.61)	18.12 (1.71)	18.70 (2.05)	19.17 (2.24)	20.14 (2.40)	22.23 (3.85)	26.39 (5.45)
-1.0	15.78 (1.57)	15.54 (1.65)	15.82 (1.63)	15.90 (1.80)	16.58 (2.16)	16.99 (2.40)	17.88 (2.81)	19.85 (3.72)	24.79 (5.63)
-0.5	13.35 (1.38)	13.45 (1.45)	13.54 (1.60)	13.89 (1.93)	14.58 (1.84)	15.16 (2.25)	15.58 (2.68)	16.88 (3.22)	18.97 (3.72)
0	12.43 (0.86)	12.52 (1.05)	12.65 (1.08)	12.88 (1.16)	12.99 (1.18)	13.52 (1.31)	13.85 (1.34)	15.12 (1.68)	16.85 (2.31)
0.5	13.62 (1.62)	13.48 (1.56)	13.82 (1.55)	13.91 (1.56)	14.13 (1.50)	14.42 (1.36)	15.05 (1.45)	16.24 (2.40)	18.13 (2.96)
1.0	15.74 (1.54)	15.75 (1.55)	15.69 (1.52)	15.94 (1.56)	16.23 (1.91)	16.71 (2.34)	17.01 (2.29)	19.13 (3.54)	21.60 (5.31)
1.5	17.27 (1.64)	17.40 (1.39)	17.59 (1.69)	17.74 (1.62)	18.18 (2.14)	18.88 (2.82)	19.75 (2.73)	21.28 (3.44)	26.37 (6.48)
2.0	18.86 (1.33)	18.89 (1.22)	19.12 (1.49)	20.03 (1.89)	20.34 (1.70)	20.88 (2.34)	21.38 (2.44)	25.30 (4.91)	30.25 (6.38)
2.5	19.66 (1.06)	20.04 (1.32)	20.59 (1.76)	21.05 (1.90)	21.73 (2.00)	22.60 (2.37)	24.37 (3.61)	27.26 (5.24)	33.29 (6.04)

Entries include average # of test items (n) and corresponding standard deviations listed below in parentheses.

Average number of test items (n) for true theta (θ) and corresponding s.d. with 3 initial test items s.e.(θ) of 0.5

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	10.64 (1.20)	10.30 (0.87)	10.23 (0.95)	10.24 (0.91)	10.10 (0.93)	10.47 (1.17)	10.72 (1.41)	11.91 (2.12)	13.16 (3.02)
-2.0	9.66 (1.09)	9.54 (0.96)	9.43 (1.88)	9.43 (0.84)	9.50 (0.97)	9.86 (1.04)	10.29 (1.31)	11.34 (1.94)	12.73 (2.96)
-1.5	8.76 (1.01)	8.57 (0.97)	8.87 (1.00)	8.89 (0.96)	8.92 (1.16)	9.24 (1.08)	9.96 (1.37)	10.74 (1.88)	12.30 (3.13)
-1.0	7.81 (1.00)	7.91 (1.00)	8.07 (1.02)	8.18 (1.02)	8.27 (1.03)	8.78 (1.18)	8.87 (1.51)	10.41 (2.18)	11.67 (2.67)
-0.5	7.01 (0.87)	6.92 (0.84)	7.16 (1.02)	7.42 (1.02)	7.37 (1.00)	7.86 (1.26)	8.14 (1.36)	8.97 (1.95)	10.21 (2.58)
0	6.80 (0.65)	6.71 (0.76)	6.77 (0.75)	6.98 (0.85)	7.06 (0.74)	7.28 (0.82)	7.66 (1.19)	8.48 (1.37)	9.49 (2.36)
0.5	7.17 (0.88)	7.18 (0.86)	7.25 (0.86)	7.34 (0.97)	7.25 (0.94)	7.49 (0.89)	7.55 (0.76)	8.18 (1.29)	9.46 (2.10)
1.0	7.82 (0.88)	7.82 (0.95)	7.90 (1.07)	8.10 (1.09)	8.10 (1.09)	8.32 (1.14)	8.48 (1.31)	8.88 (1.65)	10.84 (2.95)
1.5	8.75 (1.10)	9.00 (1.29)	8.98 (1.33)	9.50 (1.42)	9.42 (1.56)	9.59 (1.55)	9.87 (1.67)	11.14 (2.44)	12.61 (3.32)
2.0	9.76 (1.11)	10.08 (0.97)	10.22 (1.16)	10.34 (1.16)	10.85 (1.50)	11.34 (1.97)	11.46 (2.27)	12.99 (2.70)	16.25 (4.56)
2.5	10.66 (0.96)	10.92 (1.14)	11.18 (1.23)	11.39 (1.37)	11.55 (1.46)	12.29 (1.71)	12.87 (2.17)	15.52 (3.48)	19.53 (3.85)

Entries include average # of test items (n) and corresponding standard deviations listed below in parentheses.

Average number of test items (n) for true theta (θ) and corresponding s.d. with 5 initial test items s.e.(θ) of 0.5

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	12.10 (1.11)	11.96 (0.89)	11.85 (0.97)	11.83 (0.94)	11.82 (0.94)	12.04 (1.10)	12.58 (1.26)	13.43 (2.12)	15.23 (2.89)
-2.0	11.13 (1.40)	10.87 (1.13)	11.08 (1.10)	11.10 (1.24)	11.06 (1.18)	11.42 (1.42)	11.87 (1.62)	12.96 (2.01)	14.66 (3.15)
-1.5	9.93 (1.39)	9.93 (1.13)	9.96 (1.31)	10.06 (1.26)	10.41 (1.40)	10.57 (1.62)	11.03 (1.60)	12.02 (2.58)	13.98 (3.24)
-1.0	8.48 (1.36)	8.64 (1.49)	8.94 (1.55)	8.86 (1.73)	9.07 (1.71)	9.48 (2.16)	10.14 (2.20)	10.32 (2.52)	11.87 (3.72)
-0.5	7.36 (1.19)	7.45 (1.17)	7.57 (1.25)	7.48 (1.29)	7.60 (1.50)	7.47 (1.65)	8.05 (1.82)	8.17 (1.95)	9.07 (2.48)
0	6.71 (0.87)	6.77 (0.91)	6.82 (0.88)	6.81 (0.80)	6.95 (0.89)	6.85 (0.72)	6.97 (1.10)	7.47 (1.32)	7.75 (1.68)
0.5	7.30 (1.09)	7.45 (1.11)	7.46 (1.13)	7.36 (1.12)	7.31 (1.05)	7.17 (1.01)	7.36 (1.19)	7.88 (1.34)	8.41 (1.56)
1.0	8.81 (1.38)	8.87 (1.40)	8.92 (1.53)	8.82 (1.68)	8.96 (1.62)	8.74 (1.64)	8.82 (1.68)	9.28 (1.93)	9.91 (2.76)
1.5	10.12 (1.17)	10.40 (1.15)	10.25 (1.21)	10.25 (1.35)	10.35 (1.31)	10.63 (1.77)	10.85 (1.95)	11.31 (2.70)	12.69 (3.51)
2.0	11.06 (1.27)	11.21 (1.24)	11.17 (1.39)	11.47 (1.47)	11.43 (1.80)	11.92 (1.84)	12.22 (2.14)	13.57 (3.17)	17.24 (4.18)
2.5	11.98 (1.07)	12.33 (1.17)	12.44 (1.41)	12.66 (1.39)	12.83 (1.52)	13.59 (1.78)	14.15 (2.09)	16.44 (2.81)	19.15 (3.71)

Entries include average # of test items (n) and corresponding standard deviations listed below in parentheses.

Average number of test items (n) for true theta (θ) and corresponding s.d. with 7 initial test items s.e.(θ) of 0.5

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	13.62 (1.19)	13.40 (1.01)	13.68 (0.98)	13.61 (0.94)	13.59 (0.92)	14.03 (1.01)	14.22 (1.55)	15.51 (2.16)	16.77 (2.98)
-2.0	12.67 (1.41)	12.54 (1.29)	12.60 (1.24)	12.78 (1.35)	13.08 (1.28)	13.11 (1.56)	13.58 (1.85)	14.44 (1.98)	16.00 (3.14)
-1.5	11.21 (1.57)	11.33 (1.41)	11.42 (1.37)	11.45 (1.51)	11.63 (1.64)	11.75 (1.65)	12.36 (2.12)	13.13 (2.42)	14.79 (3.83)
-1.0	9.82 (1.53)	10.03 (1.57)	9.87 (1.45)	10.00 (1.66)	10.01 (1.69)	10.30 (1.92)	10.11 (2.12)	10.91 (2.61)	12.12 (3.29)
-0.5	8.47 (0.88)	8.47 (0.81)	8.45 (0.88)	8.45 (0.88)	8.52 (1.01)	8.63 (1.06)	8.60 (1.15)	9.22 (1.87)	9.11 (2.16)
0	8.17 (0.53)	8.17 (0.57)	8.16 (0.53)	8.10 (0.39)	8.16 (0.56)	8.17 (0.55)	8.17 (0.53)	8.30 (0.92)	8.15 (0.64)
0.5	8.62 (1.05)	8.48 (1.03)	8.49 (1.10)	8.62 (1.20)	8.44 (0.95)	8.45 (0.97)	8.39 (0.85)	8.62 (1.20)	8.61 (1.33)
1.0	9.99 (1.48)	9.70 (1.64)	9.91 (1.63)	9.52 (1.64)	9.69 (1.57)	9.79 (1.58)	9.71 (1.73)	9.75 (2.01)	10.61 (2.77)
1.5	10.98 (1.60)	10.96 (1.93)	10.89 (1.92)	10.91 (1.92)	11.14 (2.04)	11.22 (2.13)	11.05 (2.22)	12.00 (2.67)	13.39 (3.85)
2.0	12.73 (1.35)	12.61 (1.58)	12.72 (1.54)	12.74 (1.68)	13.01 (1.76)	13.22 (1.87)	13.60 (2.39)	15.07 (3.12)	17.62 (4.25)
2.5	13.92 (1.20)	14.01 (1.30)	14.09 (1.32)	14.17 (1.46)	14.61 (1.76)	15.21 (1.98)	15.79 (2.28)	17.48 (2.76)	20.32 (3.79)

Table entries include average # of test items (n) and corresponding standard deviations listed below in parentheses.

APPENDIX CAverage theta estimates ($\hat{\theta}$) for true theta (θ) and corresponding s.d. with 3 initial test items s.e(θ) of 0.2

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	-2.46 (0.21)	-2.48 (0.23)	-2.50 (0.23)	-2.50 (0.20)	-2.53 (0.19)	-2.51 (0.21)	-2.52 (0.22)	-2.50 (0.21)	-2.49 (0.18)
-2.0	-1.98 (0.21)	-1.98 (0.19)	-1.98 (0.19)	-2.00 (0.18)	-1.97 (0.17)	-1.98 (0.17)	-1.96 (0.21)	-2.01 (0.19)	-2.01 (0.19)
-1.5	-1.51 (0.19)	-1.50 (0.17)	-1.49 (0.19)	-1.50 (0.17)	-1.49 (0.19)	-1.49 (0.18)	-1.51 (0.19)	-1.47 (0.19)	-1.53 (0.19)
-1.0	-1.00 (0.20)	-1.00 (0.21)	-0.99 (0.20)	-0.97 (0.19)	-0.99 (0.19)	-1.00 (0.20)	-1.02 (0.22)	-1.03 (0.20)	-1.00 (0.19)
-0.5	-0.49 (0.20)	-0.50 (0.19)	-0.49 (0.20)	-0.50 (0.19)	-0.52 (0.18)	-0.51 (0.21)	-0.53 (0.22)	-0.49 (0.20)	-0.53 (0.22)
0	0.00 (0.21)	0.00 (0.21)	-0.01 (0.22)	-0.02 (0.20)	0.00 (0.21)	0.00 (0.20)	0.01 (0.20)	-0.03 (0.23)	-0.03 (0.19)
0.5	0.54 (0.18)	0.53 (0.20)	0.53 (0.19)	0.53 (0.21)	0.48 (0.19)	0.47 (0.19)	0.46 (0.18)	0.46 (0.20)	0.47 (0.24)
1.0	0.99 (0.18)	0.97 (0.18)	0.97 (0.19)	0.97 (0.20)	1.00 (0.20)	0.98 (0.23)	0.96 (0.26)	0.94 (0.21)	1.00 (0.21)
1.5	1.53 (0.20)	1.50 (0.21)	1.50 (0.23)	1.50 (0.24)	1.48 (0.22)	1.48 (0.20)	1.47 (0.19)	1.46 (0.23)	1.48 (0.21)
2.0	1.98 (0.22)	1.98 (0.22)	1.95 (0.22)	1.97 (0.21)	1.99 (0.20)	1.97 (0.19)	1.96 (0.22)	1.98 (0.19)	1.97 (0.20)
2.5	2.54 (0.21)	2.52 (0.17)	2.52 (0.20)	2.49 (0.22)	2.47 (0.23)	2.50 (0.20)	2.50 (0.20)	2.46 (0.20)	2.48 (0.19)

Table entries include average ability estimates ($\hat{\theta}$) and corresponding standard deviations listed below in parentheses.

Average theta estimates ($\hat{\theta}$) for true theta (θ) and corresponding s.d. with 5 initial test items s.e(θ) of 0.2

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	-2.50 (0.22)	-2.48 (0.22)	-2.51 (0.20)	-2.53 (0.19)	-2.49 (0.22)	-2.49 (0.20)	-2.52 (0.22)	-2.50 (0.21)	-2.49 (0.21)
-2.0	-1.98 (0.20)	-1.99 (0.19)	-1.98 (0.19)	-2.00 (0.19)	-1.99 (0.20)	-2.00 (0.20)	-1.99 (0.19)	-1.98 (0.19)	-2.00 (0.20)
-1.5	-1.51 (0.20)	-1.51 (0.19)	-1.49 (0.19)	-1.49 (0.20)	-1.48 (0.20)	-1.50 (0.19)	-1.49 (0.19)	-1.50 (0.20)	-1.53 (0.21)
-1.0	-0.99 (0.20)	-0.97 (0.21)	-0.99 (0.21)	-0.99 (0.20)	-1.00 (0.18)	-1.01 (0.23)	-1.01 (0.19)	-1.00 (0.21)	-0.98 (0.20)
-0.5	-0.50 (0.18)	-0.49 (0.19)	-0.51 (0.18)	-0.49 (0.18)	-0.51 (0.21)	-0.51 (0.19)	-0.49 (0.23)	-0.53 (0.20)	-0.56 (0.22)
0	0.01 (0.19)	0.02 (0.20)	-0.02 (0.21)	0.01 (0.21)	0.01 (0.21)	0.00 (0.19)	-0.04 (0.19)	-0.04 (0.20)	-0.03 (0.20)
0.5	0.53 (0.21)	0.51 (0.20)	0.52 (0.19)	0.49 (0.21)	0.48 (0.18)	0.46 (0.19)	0.48 (0.20)	0.46 (0.22)	0.47 (0.21)
1.0	0.97 (0.19)	0.99 (0.19)	0.99 (0.19)	0.99 (0.20)	0.99 (0.20)	1.00 (0.22)	0.96 (0.20)	0.97 (0.22)	0.98 (0.20)
1.5	1.49 (0.21)	1.49 (0.22)	1.48 (0.23)	1.51 (0.24)	1.46 (0.21)	1.47 (0.19)	1.45 (0.21)	1.48 (0.21)	1.47 (0.18)
2.0	1.98 (0.20)	1.99 (0.21)	2.01 (0.22)	1.99 (0.21)	2.01 (0.22)	1.97 (0.20)	1.97 (0.21)	1.98 (0.21)	1.98 (0.20)
2.5	2.53 (0.20)	2.54 (0.21)	2.51 (0.21)	2.48 (0.23)	2.49 (0.22)	2.48 (0.22)	2.50 (0.21)	2.46 (0.20)	2.48 (0.19)

Table entries include average ability estimates ($\hat{\theta}$) and corresponding standard deviations listed below in parentheses.

Average theta estimates ($\hat{\theta}$) for true theta (θ) and corresponding s.d. with 7 initial test items and s.e(θ) of 0.2

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	-2.47 (0.24)	-2.51 (0.22)	-2.51 (0.23)	-2.49 (0.19)	-2.48 (0.21)	-2.48 (0.23)	-2.51 (0.18)	-2.51 (0.21)	-2.49 (0.22)
-2.0	-1.98 (0.21)	-1.97 (0.20)	-1.97 (0.19)	-1.97 (0.20)	-1.99 (0.21)	-2.01 (0.19)	-2.00 (0.21)	-1.98 (0.18)	-2.01 (0.19)
-1.5	-1.51 (0.20)	-1.50 (0.20)	-1.50 (0.19)	-1.49 (0.20)	-1.51 (0.18)	-1.51 (0.20)	-1.50 (0.17)	-1.49 (0.19)	-1.50 (0.22)
-1.0	-0.98 (0.19)	-0.98 (0.20)	-0.99 (0.20)	-0.98 (0.18)	-0.99 (0.20)	-1.01 (0.19)	-1.01 (0.20)	-0.98 (0.22)	-1.03 (0.19)
-0.5	-0.50 (0.20)	-0.50 (0.18)	-0.50 (0.20)	-0.50 (0.19)	-0.50 (0.20)	-0.50 (0.19)	-0.48 (0.22)	-0.52 (0.18)	-0.55 (0.25)
0	0.00 (0.20)	0.01 (0.20)	0.01 (0.17)	0.02 (0.20)	0.02 (0.22)	0.01 (0.19)	-0.03 (0.20)	-0.05 (0.20)	-0.03 (0.22)
0.5	0.50 (0.19)	0.52 (0.21)	0.48 (0.22)	0.47 (0.21)	0.49 (0.19)	0.46 (0.19)	0.50 (0.23)	0.46 (0.20)	0.48 (0.20)
1.0	0.99 (0.19)	0.99 (0.19)	0.99 (0.20)	0.99 (0.21)	0.98 (0.22)	0.98 (0.22)	0.94 (0.22)	0.98 (0.22)	0.98 (0.20)
1.5	1.52 (0.21)	1.50 (0.21)	1.51 (0.22)	1.50 (0.22)	1.46 (0.22)	1.49 (0.21)	1.45 (0.20)	1.45 (0.22)	1.47 (0.21)
2.0	1.98 (0.21)	1.98 (0.22)	2.00 (0.19)	2.00 (0.19)	1.99 (0.21)	1.98 (0.21)	1.97 (0.20)	1.98 (0.22)	1.98 (0.19)
2.5	2.50 (0.20)	2.50 (0.21)	2.47 (0.21)	2.46 (0.22)	2.47 (0.20)	2.49 (0.20)	2.48 (0.20)	2.46 (0.19)	2.49 (0.22)

Table entries include average ability estimates ($\hat{\theta}$) and corresponding standard deviations listed below in parentheses.

Average theta estimates ($\hat{\theta}$) for true theta (θ) and corresponding s.d. with 3 initial test items s.e(θ) of 0.35

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	-2.46 (0.35)	-2.52 (0.40)	-2.50 (0.33)	-2.53 (0.32)	-2.58 (0.33)	-2.58 (0.33)	-2.55 (0.33)	-2.57 (0.32)	-2.56 (0.31)
-2.0	-2.02 (0.37)	-1.95 (0.39)	-2.03 (0.38)	-1.99 (0.36)	-2.02 (0.38)	-2.00 (0.33)	-2.05 (0.38)	-2.04 (0.39)	-2.05 (0.38)
-1.5	-1.51 (0.36)	-1.51 (0.34)	-1.49 (0.33)	-1.53 (0.37)	-1.52 (0.36)	-1.48 (0.35)	-1.49 (0.36)	-1.53 (0.37)	-1.47 (0.35)
-1.0	-1.01 (0.39)	-1.04 (0.32)	-1.00 (0.37)	-1.00 (0.38)	-1.02 (0.35)	-1.03 (0.32)	-0.99 (0.39)	-0.94 (0.34)	-1.07 (0.36)
-0.5	-0.46 (0.33)	-0.43 (0.33)	-0.48 (0.39)	-0.49 (0.35)	-0.47 (0.32)	-0.48 (0.32)	-0.49 (0.29)	-0.56 (0.31)	-0.49 (0.37)
0	0.02 (0.36)	0.01 (0.34)	0.02 (0.35)	-0.01 (0.30)	-0.02 (0.36)	-0.02 (0.30)	-0.05 (0.35)	0.00 (0.32)	-0.05 (0.31)
0.5	0.51 (0.36)	0.47 (0.35)	0.47 (0.32)	0.49 (0.34)	0.51 (0.34)	0.45 (0.34)	0.51 (0.35)	0.45 (0.33)	0.49 (0.36)
1.0	1.02 (0.37)	1.02 (0.34)	1.01 (0.36)	1.02 (0.33)	1.03 (0.34)	1.03 (0.30)	0.98 (0.34)	0.99 (0.37)	0.97 (0.38)
1.5	1.54 (0.32)	1.54 (0.32)	1.52 (0.31)	1.45 (0.33)	1.44 (0.36)	1.41 (0.38)	1.44 (0.40)	1.45 (0.41)	1.41 (0.35)
2.0	1.94 (0.40)	1.96 (0.35)	1.97 (0.36)	1.95 (0.34)	2.00 (0.37)	1.98 (0.38)	1.93 (0.35)	1.95 (0.36)	1.93 (0.36)
2.5	2.53 (0.36)	2.54 (0.33)	2.52 (0.34)	2.54 (0.35)	2.47 (0.35)	2.42 (0.35)	2.50 (0.38)	2.46 (0.36)	2.47 (0.38)

Table entries include average ability estimates ($\hat{\theta}$) and corresponding standard deviations listed below in parentheses.

Average theta estimates ($\hat{\theta}$) for true theta (θ) and corresponding s.d. with 5 initial test items s.e(θ) of 0.35

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	-2.51 (0.37)	-2.55 (0.30)	-2.53 (0.33)	-2.55 (0.32)	-2.53 (0.31)	-2.58 (0.38)	-2.53 (0.35)	-2.59 (0.33)	-2.51 (0.37)
-2.0	-1.97 (0.37)	-1.96 (0.40)	-1.99 (0.39)	-2.00 (0.33)	-2.00 (0.37)	-2.05 (0.36)	-2.04 (0.34)	-2.05 (0.36)	-2.10 (0.39)
-1.5	-1.46 (0.37)	-1.49 (0.34)	-1.43 (0.35)	-1.47 (0.38)	-1.54 (0.32)	-1.55 (0.35)	-1.51 (0.39)	-1.50 (0.37)	-1.49 (0.38)
-1.0	-1.01 (0.39)	-1.01 (0.40)	-1.01 (0.39)	-0.99 (0.37)	-0.96 (0.34)	-0.93 (0.30)	-0.98 (0.34)	-0.98 (0.38)	-1.05 (0.38)
-0.5	-0.43 (0.32)	-0.46 (0.32)	-0.45 (0.33)	-0.48 (0.32)	-0.48 (0.33)	-0.54 (0.36)	-0.54 (0.32)	-0.48 (0.36)	-0.45 (0.32)
0	-0.01 (0.31)	-0.04 (0.30)	0.02 (0.30)	-0.03 (0.30)	-0.08 (0.34)	0.00 (0.36)	-0.01 (0.36)	0.04 (0.30)	-0.02 (0.31)
0.5	0.53 (0.32)	0.52 (0.35)	0.51 (0.33)	0.51 (0.35)	0.51 (0.34)	0.53 (0.32)	0.50 (0.30)	0.47 (0.32)	0.52 (0.30)
1.0	1.03 (0.36)	1.01 (0.37)	1.06 (0.32)	1.03 (0.36)	1.02 (0.32)	0.98 (0.32)	0.96 (0.34)	0.98 (0.38)	0.99 (0.36)
1.5	1.51 (0.34)	1.48 (0.32)	1.50 (0.32)	1.48 (0.38)	1.45 (0.41)	1.42 (0.39)	1.48 (0.39)	1.48 (0.37)	1.39 (0.34)
2.0	1.96 (0.36)	1.95 (0.36)	1.93 (0.36)	1.96 (0.36)	2.00 (0.35)	2.00 (0.36)	1.97 (0.35)	1.96 (0.34)	1.92 (0.35)
2.5	2.54 (0.38)	2.52 (0.37)	2.50 (0.35)	2.49 (0.35)	2.50 (0.31)	2.54 (0.37)	2.47 (0.35)	2.41 (0.33)	2.50 (0.38)

Table entries include average ability estimates ($\hat{\theta}$) and corresponding standard deviations listed below in parentheses.

Average theta estimates ($\hat{\theta}$) for true theta (θ) and corresponding s.d. with 7 initial test items s.e(θ) of 0.35

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	-2.49 (0.36)	-2.52 (0.33)	-2.53 (0.35)	-2.53 (0.35)	-2.52 (0.39)	-2.55 (0.30)	-2.54 (0.34)	-2.53 (0.34)	-2.56 (0.36)
-2.0	-1.97 (0.37)	-2.00 (0.38)	-1.97 (0.42)	-1.99 (0.35)	-2.00 (0.37)	-2.02 (0.35)	-1.97 (0.33)	-2.03 (0.39)	-2.02 (0.37)
-1.5	-1.46 (0.34)	-1.49 (0.37)	-1.49 (0.34)	-1.50 (0.34)	-1.52 (0.37)	-1.49 (0.37)	-1.53 (0.37)	-1.49 (0.36)	-1.50 (0.37)
-1.0	-1.01 (0.34)	-0.98 (0.36)	-0.99 (0.36)	-0.90 (0.31)	-0.96 (0.33)	-0.93 (0.35)	-0.95 (0.38)	-1.01 (0.32)	-1.00 (0.36)
-0.5	-0.48 (0.35)	-0.48 (0.35)	-0.48 (0.34)	-0.47 (0.38)	-0.53 (0.35)	-0.54 (0.34)	-0.57 (0.36)	-0.51 (0.34)	-0.52 (0.31)
0	0.00 (0.32)	-0.02 (0.33)	0.01 (0.35)	0.01 (0.33)	0.01 (0.34)	0.00 (0.38)	0.04 (0.35)	0.01 (0.33)	-0.03 (0.30)
0.5	0.54 (0.35)	0.51 (0.33)	0.53 (0.35)	0.54 (0.35)	0.52 (0.34)	0.51 (0.32)	0.54 (0.28)	0.44 (0.34)	0.50 (0.30)
1.0	1.01 (0.31)	1.04 (0.35)	1.01 (0.34)	1.00 (0.35)	1.00 (0.36)	0.95 (0.33)	0.93 (0.33)	0.98 (0.35)	0.95 (0.38)
1.5	1.43 (0.38)	1.45 (0.34)	1.44 (0.34)	1.43 (0.39)	1.45 (0.39)	1.44 (0.39)	1.50 (0.35)	1.46 (0.38)	1.41 (0.37)
2.0	2.01 (0.37)	2.00 (0.38)	2.01 (0.40)	2.06 (0.34)	2.04 (0.34)	2.02 (0.31)	1.97 (0.39)	1.97 (0.37)	1.97 (0.36)
2.5	2.48 (0.40)	2.50 (0.34)	2.46 (0.39)	2.49 (0.37)	2.54 (0.38)	2.48 (0.33)	2.48 (0.35)	2.45 (0.37)	2.47 (0.34)

Table entries include average ability estimates ($\hat{\theta}$) and corresponding standard deviations listed below in parentheses.

Average theta estimates ($\hat{\theta}$) for true theta (θ) and corresponding s.d. with 3 initial test items s.e(θ) of 0.5

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	-2.50 (0.46)	-2.50 (0.49)	-2.44 (0.47)	-2.50 (0.44)	-2.47 (0.49)	-2.55 (0.43)	-2.64 (0.49)	-2.53 (0.47)	-2.59 (0.47)
-2.0	-1.97 (0.49)	-2.03 (0.47)	-1.92 (0.49)	-1.93 (0.46)	-2.01 (0.47)	-2.05 (0.50)	-2.06 (0.49)	-2.12 (0.55)	-2.06 (0.51)
-1.5	-1.52 (0.52)	-1.45 (0.49)	-1.46 (0.50)	-1.47 (0.51)	-1.52 (0.53)	-1.49 (0.44)	-1.64 (0.54)	-1.60 (0.49)	-1.44 (0.47)
-1.0	-0.95 (0.51)	-1.00 (0.51)	-0.99 (0.50)	-1.00 (0.45)	1.02 (0.47)	-1.05 (0.44)	-0.95 (0.52)	-1.05 (0.47)	-1.06 (0.50)
-0.5	-0.42 (0.57)	-0.39 (0.52)	-0.40 (0.47)	-0.45 (0.59)	-0.44 (0.55)	-0.48 (0.48)	-0.51 (0.51)	-0.57 (0.46)	-0.58 (0.54)
0	-0.01 (0.52)	-0.02 (0.50)	0.00 (0.46)	0.00 (0.47)	-0.01 (0.50)	0.05 (0.44)	-0.07 (0.54)	-0.08 (0.48)	-0.04 (0.46)
0.5	0.58 (0.45)	0.49 (0.48)	0.54 (0.49)	0.51 (0.44)	0.47 (0.45)	0.51 (0.49)	0.48 (0.50)	0.49 (0.47)	0.49 (0.39)
1.0	0.95 (0.44)	0.90 (0.47)	0.94 (0.50)	0.90 (0.50)	0.91 (0.51)	0.98 (0.51)	0.96 (0.53)	0.96 (0.44)	0.92 (0.45)
1.5	1.46 (0.56)	1.43 (0.56)	1.43 (0.54)	1.52 (0.55)	1.49 (0.56)	1.44 (0.48)	1.50 (0.41)	1.46 (0.49)	1.31 (0.52)
2.0	2.04 (0.48)	2.07 (0.41)	2.04 (0.46)	1.93 (0.41)	2.03 (0.48)	2.00 (0.54)	1.90 (0.55)	1.88 (0.47)	1.85 (0.55)
2.5	2.48 (0.43)	2.46 (0.47)	2.44 (0.43)	2.39 (0.50)	2.35 (0.45)	2.38 (0.48)	2.33 (0.51)	2.35 (0.53)	2.45 (0.51)

Table entries include average ability estimates ($\hat{\theta}$) and corresponding standard deviations listed below in parentheses.

Average theta estimates ($\hat{\theta}$) for true theta (θ) and corresponding s.d. with 5 initial test items s.e(θ) of 0.5

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	-2.52 (0.53)	-2.59 (0.40)	-2.51 (0.47)	-2.47 (0.48)	-2.51 (0.45)	-2.54 (0.47)	-2.58 (0.50)	-2.65 (0.40)	-2.60 (0.48)
-2.0	-1.97 (0.55)	-1.96 (0.52)	-1.95 (0.50)	-1.99 (0.52)	-1.98 (0.59)	-2.00 (0.51)	-2.02 (0.52)	-2.09 (0.49)	-2.10 (0.51)
-1.5	-1.46 (0.51)	-1.49 (0.48)	-1.45 (0.47)	-1.48 (0.51)	-1.52 (0.53)	-1.54 (0.58)	-1.47 (0.52)	-1.46 (0.54)	-1.53 (0.49)
-1.0	-0.91 (0.46)	-0.96 (0.55)	-1.01 (0.52)	-0.98 (0.56)	-0.95 (0.46)	-0.97 (0.50)	-1.07 (0.53)	-1.00 (0.54)	-1.04 (0.49)
-0.5	-0.43 (0.47)	-0.49 (0.45)	-0.50 (0.45)	-0.49 (0.47)	-0.49 (0.47)	-0.49 (0.43)	-0.52 (0.50)	-0.49 (0.43)	-0.50 (0.49)
0	0.03 (0.43)	0.07 (0.43)	-0.05 (0.44)	0.02 (0.41)	0.02 (0.47)	0.02 (0.46)	0.01 (0.43)	-0.01 (0.46)	0.01 (0.49)
0.5	0.44 (0.45)	0.46 (0.46)	0.45 (0.47)	0.41 (0.46)	0.42 (0.48)	0.37 (0.45)	0.46 (0.45)	0.45 (0.49)	0.49 (0.41)
1.0	1.01 (0.51)	1.00 (0.54)	1.01 (0.51)	1.01 (0.54)	1.00 (0.55)	0.97 (0.50)	0.99 (0.49)	1.00 (0.43)	0.87 (0.43)
1.5	1.55 (0.45)	1.57 (0.44)	1.49 (0.41)	1.48 (0.41)	1.51 (0.41)	1.50 (0.47)	1.54 (0.48)	1.36 (0.48)	1.32 (0.54)
2.0	2.00 (0.53)	1.93 (0.47)	1.90 (0.51)	1.91 (0.50)	1.88 (0.53)	1.92 (0.49)	1.93 (0.50)	1.86 (0.55)	1.99 (0.48)
2.5	2.48 (0.49)	2.50 (0.52)	2.44 (0.54)	2.40 (0.49)	2.38 (0.49)	2.41 (0.51)	2.40 (0.51)	2.47 (0.47)	2.38 (0.49)

Table entries include average ability estimates ($\hat{\theta}$) and corresponding standard deviations listed below in parentheses.

Average theta estimates ($\hat{\theta}$) for true theta (θ) and corresponding s.d. with 7 initial test items s.e(θ) of 0.5

θ	Delta (δ)								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
-2.5	-2.50 (0.48)	-2.49 (0.52)	-2.61 (0.48)	-2.58 (0.52)	-2.50 (0.47)	-2.60 (0.51)	-2.51 (0.54)	-2.54 (0.52)	-2.65 (0.44)
-2.0	-2.04 (0.54)	-2.03 (0.55)	-1.98 (0.47)	-2.05 (0.55)	-2.08 (0.52)	-2.01 (0.57)	-2.05 (0.58)	-2.07 (0.56)	-2.06 (0.56)
-1.5	-1.43 (0.48)	-1.49 (0.47)	-1.50 (0.48)	-1.46 (0.47)	-1.47 (0.45)	-1.42 (0.47)	-1.51 (0.56)	-1.48 (0.50)	-1.51 (0.59)
-1.0	-0.99 (0.45)	-1.03 (0.52)	-0.97 (0.45)	-1.01 (0.45)	-0.99 (0.48)	-0.99 (0.47)	-0.95 (0.53)	-0.97 (0.50)	-1.06 (0.45)
-0.5	-0.44 (0.43)	-0.45 (0.39)	-0.44 (0.41)	-0.45 (0.42)	-0.47 (0.44)	-0.44 (0.43)	-0.44 (0.43)	-0.49 (0.47)	-0.48 (0.41)
0	-0.05 (0.49)	-0.04 (0.44)	-0.06 (0.43)	-0.06 (0.41)	-0.03 (0.44)	-0.01 (0.46)	0.02 (0.47)	-0.01 (0.47)	0.08 (0.42)
0.5	0.55 (0.46)	0.52 (0.43)	0.54 (0.43)	0.59 (0.49)	0.54 (0.43)	0.50 (0.46)	0.48 (0.43)	0.55 (0.38)	0.50 (0.37)
1.0	1.06 (0.45)	1.03 (0.41)	1.03 (0.41)	1.00 (0.40)	1.02 (0.41)	1.03 (0.40)	1.01 (0.40)	0.91 (0.42)	0.90 (0.39)
1.5	1.39 (0.43)	1.40 (0.50)	1.38 (0.51)	1.42 (0.50)	1.46 (0.50)	1.42 (0.50)	1.35 (0.51)	1.36 (0.50)	1.36 (0.53)
2.0	1.99 (0.50)	1.94 (0.53)	1.95 (0.53)	1.90 (0.50)	1.96 (0.53)	1.95 (0.48)	1.91 (0.50)	1.90 (0.47)	1.96 (0.51)
2.5	2.59 (0.49)	2.51 (0.51)	2.44 (0.45)	2.40 (0.46)	2.49 (0.50)	2.49 (0.49)	2.50 (0.47)	2.48 (0.48)	2.42 (0.52)

Table entries include average ability estimates ($\hat{\theta}$) and corresponding standard deviations listed below in parentheses.

REFERENCES

- Ackerman, T., Spray, J., Reckase, M., & Carlson, J. (1989). A comparison of the effects of random versus fixed order of item presentation via the computer. ACT Research Report Series, 89(1), 1-34.
- Allison, D. E. (1984). The effect of item-difficulty sequence, intelligence, and sex on test performance, reliability, and item difficulty and discrimination. Measurement and Evaluation in Guidance, 16(4), 211-217.
- Andrich, D. (1995). Review of Computerized adaptive testing: A primer. Psychometrika, 60(4), 615-620.
- Bergstrom, B & Lunz, M. (1999). CAT for certification and licensure. In F. Drasgow & J. Olson-Buchanan (Eds.) *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Earlbaum.
- Bergstrom, B. A., Lunz, M. E. & Gershon, R. C. (1992). Altering the level of Difficulty in computer adaptive testing. Applied Measurement in Education, 5(2), 137-149.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.) *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Crocker, L. & Algina, J. (1986). *Introduction to classical & Modern test theory*. Fort Worth: Harcourt Brace Jovanovich.
- Green, B. (1983). Adaptive testing by computer. In R. B. Ekstrom (Ed.) *Measurement, Technology, and Individuality in Education*. San Francisco: Jossey-Bass.
- Guernsey, L. (2000). "An ever-changing course: Taking admissions tests on computer." New York Times, (6 August 2000): Education Life, 32-33.
- Guilford, J. P. (1954). *Psychometric Methods*. New York: McGraw Hill.
- Halkitis, P. & Leahy, J. (1993). Computerized adaptive testing: The future is upon us. Nursing and Health Care, 14(7), 379-385.
- Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer – Nijhoff.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

- Hambleton, R. K. & Traub, R. E. (1974). The effects of item order on test performance and stress. The Journal of Experimental Education, 43(1), 40-46.
- Jacobson, R. L. (1993). New computer technique seen producing a revolution in educational testing. Chronicle of Higher Education, 40(4), A22-23,26.
- Kingsbury, G. G. & Houser, R. L. (1999). Developing computerized adaptive tests for school children. In F. Drasgow & J. Olson-Buchanan (Eds.) *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Earlbaum.
- Legg, S. M. & Buhr, D. C. (1992). Computerized adaptive testing with different groups. Educational Measurement: Issues and Practices, 11(2), 23-27.
- Lord, F. M. (1952). A theory of test scores. Psychometric Monographs, no.7.
- Lord, F. M. (1953). The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 13, 517-549.
- Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14(2), 117-138.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Earlbaum.
- Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lunz, M. E. & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. Journal of Educational Measurement, 31(3), 251-263.
- McBride, J. (1985). Computerized adaptive testing. Educational Leadership, 43(2), 25-28.
- McBride, J. & Martin, J. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. Weiss (Ed.) *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- Mills, C. (1999). Development and introduction of a computer adaptive Graduate Record Examinations general test. In F. Drasgow & J. Olson-Buchanan (Eds.) *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Earlbaum.

- Olson-Buchanan, J & Drasgow, F. (1999). Beyond bells and whistles: An introduction to computerized assessment. In F. Drasgow & J. Olson-Buchanan (Eds.) *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Earlbaum.
- Plake, B. S. (1995). *Guidelines for computerized-adaptive test development and use in education*. Washington: American Council on Education.
- Plake, B. S., Ansorge, C. J., Parker, C. S., & Lowry, S. R. (1982). Effects of item arrangement, knowledge of arrangement, test anxiety and sex on performance. Journal of Educational Measurement, 19(1), 49-57.
- Ponsoda, V., Olea, J., Rodriguez, M. S., & Revuelta, J. (1999). The effects of test difficulty manipulation in computerized adaptive testing and self-adaptive testing. Applied Measurement in Education, 12(2), 167-184.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded Edition, 1980, with foreword and afterword by B. D. Wright. Chicago: University of Chicago Press, 1980).
- Ree, M. & Jensen, H. (1983). Effects of sample size on linear equating of item characteristic curve parameters. In D. Weiss (Ed.) *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- Segall, D. & Moreno, K. (1999). Development of the computerized adaptive testing version of the armed services vocational aptitude battery. In F. Drasgow & J. Olson-Buchanan (Eds.) *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Earlbaum.
- Straetmans, G. & Eggen, T. (1998). Computerized Adaptive Testing: What it is and how it works. Educational Technology, 38(1), 45-52.
- Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.) *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Earlbaum.
- Thissen, D. & Mislevy, R. (1990). Testing algorithms. In H. Wainer (Ed.) *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Earlbaum.
- Tippets, E. & Benson, J. (1989). The effect of item arrangement on test anxiety. Applied Measurement in Education, 2(4), 289-296.

- Trabin, T. & Weiss, D. (1983). The person response curve: Fit of individuals to item response theory models. In D. Weiss (Ed.) *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 14(2), 181-196.
- Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. Machine-mediated Learning, 2, 271-282.
- Wainer, H. (1983). Are we correcting for guessing in the wrong direction? In D. Weiss (Ed.) *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Earlbaum.
- Wainer, H & Mislevy, R. (1990). Item response theory, item calibration and proficiency estimation. In H. Wainer (Ed.) *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Earlbaum.
- Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- 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.
- Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.
- Zickar, M., Overton, R., Taylor, L., & Harms, H. (1999). The development of a computerized selection system for computer programmers in a financial services company. In F. Drasgow & J. Olson-Buchanan (Eds.) *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Earlbaum.