UNIVERSITY OF ALBERTA

THE CONSTRUCTION AND EVALUATION OF AUTOMATED PARALLEL FORMS AND MULTIPLE PARALLEL PANELS IN COMPUTER-ADAPTIVE SEQUENTIAL TESTING

By

Keith Andrew Boughton ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Educational Psychology

Edmonton, Alberta

Fall, 2001

0-612-68912-3

University of Alberta
Library Release Form

Name of Author: KEITH ANDREW BOUGHTON

Title of Thesis: THE CONSTRUCTION AND EVALUATION OF AUTOMATED PARALLEL FORMS AND MULTIPLE PARALLEL PANELS IN COMPUTER-ADAPTIVE SEQUENTIAL TESTING

Degree: DOCTOR OF PHILOSOPHY

Year Degree Granted: 2001

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

September 4, 2001

Keith A. Boughton
#103C 15111 45 Ave.
Edmonton, Alberta.
T6H 5K8

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled THE CONSTRUCTION AND EVALUATION OF AUTOMATED PARALLEL FORMS AND MULTIPLE PARALLEL PANELS IN COMPUTER-ADAPTIVE SEQUENTIAL TESTING submitted by KEITH ANDREW BOUGHTON in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dr. Mark J. Gierl

Dr. W. Todd Rogers

Michael Carbonaro
Dr. Mike Carbonaro

S. Hunka
Dr. Steve M. Hunka

Dr. Stephen Norris

Dr. Terry A. Ackerman

Date: August 31, 2001

Dedication

To my mother, Nancy Veronica Boughton.

Your beautiful spirit affects all that I achieve.

Abstract

This study investigated automated parallel forms construction using the computer-adaptive sequential testing (CAST) model for a criterion-referenced achievement test and a criterion-referenced licensure examination. The construction of the achievement examination was implemented in order to demonstrate the use and utility of the CAST model for the purpose of parallel forms construction. The construction of the licensure examination was used to demonstrate the use and utility of the CAST model for multiple parallel panels construction. There were five constraints incorporated into the production of parallel forms and panels: content area, cognitive level, test length, item exposure, and statistical targets. In order to assess the parallelism of each form, both current and new methods developed in this dissertation for evaluating parallel forms were assessed. Finally, there was an overview of item banking procedures and practical issues in the development and maintenance of item banks for the purpose of ATA parallel forms and CAST parallel panels.

After the test forms from the achievement item bank and the panels from the licensure item bank were assembled, parallelism was determined. In regards to the criterion-referenced Alberta Achievement examinations, the two ATA forms developed by the CASTISEL computer program were parallel. The handcrafted Alberta Achievement examination varied greatly across years (from1995 through 1999) with none being parallel to the target. Of the two Medical Council Panel assemblies, the First MCC panel assembly resulted in the first five stages all matching their respective alternate forms with stage six missing the target. Thus, the Second MCC panel assembly was implemented by lowering the amount of target information across the modules. This lowering of the information function targets allowed the sixth stage modules to come closer to their targets, although they did not actually meet the ratio

difference criterion for parallelism of less than 0.05. However, in consideration of the amount of information differences at the module level, the Second MCC assembly was considered parallel when using all the methods to assess parallelism considered in this study.

Acknowledgements

I wish to thank each of the following people whose help and guidance has made the completion of this dissertation possible.

To Mark Gierl, my dissertation supervisor, for your inspirational drive toward excellence. I appreciated your enthusiasm for good research and your availability and flexibility throughout my entire program. I am thankful and was fortunate to have such a supportive mentor.

To Todd Rogers, director of CRAME, for your interest, knowledge, and insightful suggestions throughout all of my endeavors. I appreciated your work ethic, the challenges you presented me with, and your unconditional assistance, both personally and academically.

To Steve Hunka, as a member of my dissertation committee, for your personal attention to detail and for ensuring that the quality of my dissertation met your high standards.

To Mike Carbonaro, Stephen Norris, and Terry Ackerman, for your interest and insightful suggestions as members of my dissertation committee.

To Fernando Cartwright, a friend and colleague, for your help in developing the numerical procedures for the assessment of parallelism found in this dissertation. I truly could not have succeeded or completed all that I set out to do without your help.

To my family: Dad, Kathy, Tracy, Jason, Danny, Wendy, Nissan, and Becky. Your curiosity, concern, and love have been deeply appreciated.

To Suzy White, partner in crime and chief editor, for your love and support at the most critical time in my program. I thank you for the countless nights of editing and revisions of my dissertation and for believing in me. I look forward to each and everyday with you by my side.

Table of Contents

List of Tables

List of Figures

Chapter I: Introduction

Computer adaptive testing (CAT) can improve the measurement process by reducing test length, improving test administration standardization, increasing measurement precision, improving testing security, increasing the flexibility for examinees by allowing for testing on-demand, and scoring and reporting results immediately (Sands & Waters, 1997; Wainer, 1990). Although van der Linden and Reese (1998) recently noted that new developments in computer technology and the acceptance of item response theory (IRT) for item bank calibration have made CAT possible, it is my contention that some testing organizations may not be ready or willing to move into a CAT framework. The benefits of a CAT system are offset by many limitations, the most severe of which is the loss of control in the assembly stage of test development by content specialists. As a result, Luecht and Nungester (1998) stated that test quality can be limited when CAT is used since this real-time automated procedure eliminates the human judgment of the item assembly and test review process. They also noted that CAT requires a great deal of trust in statistical criteria and outcomes, often at the expense of content specifications and test cohesion. Viable alternatives to a CAT system that still use item banks and statistical targets are 1) linear testing via computer-based or paper administration or 2) computer-adaptive sequential testing via computer-based administration.

Luecht and Nungester (1998) created a test assembly procedure called computer-adaptive sequential testing (CAST) that draws from the strengths of CAT while still allowing for quality assurance across test forms. A sequential test, like an item-based computer-adaptive test, could be used to reduce the number of items administered to an examinee without the loss of measurement precision, thereby reducing testing time and cost. CAST also has the added benefit of quality control because it allows the test developer to create multiple test forms that can be reviewed by content specialists before test administration. The construction of CAST modules (i.e., a set of items to be delivered together) is performed by a computer program called CASTISEL (Luecht, 1998a) that uses the normalized weighted absolute deviation heuristic (NWADH). As will be shown, this heuristic can also be used for the construction of parallel forms. The CASTISEL software is not a test delivery system; hence, issues of adaptive test administration and scoring will not be discussed here. This research is focused exclusively on the test assembly process--using CAST for both linear testing with parallel forms and computer-adaptive sequential testing with parallel “panels”.

Within the CAST framework, a test developer can generate parallel forms by constraining the first stage of the CAST assembly process while controlling for any number of variables including content and cognitive level coverage, test length, item exposure, and statistical targets across all forms. To differentiate between the automated test assembly for the purpose of parallel forms and parallel panels in computer-adaptive sequential testing, the construction of parallel forms will be referred to as “ATA” forms and the construction of parallel panels in computer-adaptive sequential testing as “CAST” forms.

Both of these methods could use automated processes to assist in the construction of parallel test forms before test administration. Whether computer-based or paper administration, the real need for automated assistance comes about as item banks grow in size, and, consequently, the parallel test development process becomes more labor intensive (Luecht, 1998b).

Purpose of Study

This study was designed to investigate the efficiency and effectiveness of computer-assisted parallel forms construction using the CAST model for a criterion-referenced achievement test and a criterion-referenced licensure examination. The construction of the achievement examination was implemented in order to demonstrate the use and utility of the CAST model for the purpose of parallel forms construction. The construction of the licensure examination was used to demonstrate the use and utility of a CAST system for multiple parallel panels construction. The following five constraints were incorporated in the production of parallel forms and panels: content area, cognitive level, test length, item exposure, and statistical targets. In order to assess the parallelism of each form, both current and new methods for evaluating parallel forms were assessed. It was found through preliminary research that the current methods for the assessment of parallelism were not satisfactory by themselves and thus new methods were developed and used with the old methods. Finally, an overview of item banking procedures and practical issues in the development and maintenance of item banks for the purpose of ATA parallel forms and CAST parallel panels will be presented.

The automation of parallel forms and parallel panels construction requires consideration of four concepts, namely: A) the nature of test score interpretation, which in the present study was criterion-referenced interpretation, B) the use of item response theory, C) the creation of the item bank, and D) the characteristics and construction of parallel forms (see Figure 1). All four considerations are reviewed in the first part of Chapter II, followed by a detailed section on computer-adaptive sequential testing.

Chapter II: Literature Review

Various developments have been made over the past 30 years in the area of achievement testing. The first development was the movement away from norm-referenced testing (NRT) towards a criterion-referenced testing (CRT) paradigm (Hambleton & Novick, 1973; see also the special issue of Applied Psychological Measurement, 1980). The second development was the movement away from classical test theory (CTT) statistics towards item response theory (IRT) statistics, at least when large samples of examinees were available for parameter estimation (Hambleton & Swaminathan, 1985). The third development was the movement away from single form construction to the development of item banks from which several forms could be constructed (van der Linden, 1986; see also the special issue of Applied Psychological Measurement, 1986). And, finally, the fourth development was the introduction of an automated test assembly process that could be used in place of or to assist with the current manual assembly process (van der Linden, 1998; see also the special issue of Applied Psychological Measurement, 1998).

A. Criterion-Referenced Testing

"A criterion-referenced test is constructed to assess the performance levels of examinees in relation to a set of well-defined objectives (or competencies)" (Hambleton, 1980, p. 421). In contrast to NRT, attempts are made in CRT to incorporate levels of achievement that (a) must be attained by students and (b) are not directly dependent on how other students perform. Thus, the quality of student performance can be interpreted in relation to the learning objectives outlined in the curriculum (Hambleton, 1980). Within a CTT framework, the scores that are obtained are interpreted in terms of a norm-referenced standard. In an IRT framework, the scores that are obtained are related to the actual items the examinee writes, and in this sense, interpretation is considered a more criterion-referenced standard (Embretson, 1996).

With regards to the Alberta Provincial achievement tests that were studied in the present research, a somewhat "eclectic" interpretive framework is applied. The framework is not strictly norm-referenced and is somewhat more criterion-referenced. The purpose of the Alberta provincial examination program is to determine how well students in the province are doing compared to what they are expected to learn. This expectation comes from pre-specified content and cognitive level blueprints that each jurisdiction across the entire province must follow. These criterion-referenced achievement tests are like many achievement tests used throughout North America and they are characterized by the need to differentiate examinees at two points on the score scale—an average cut score and a high cut score. Typically, an acceptable standard of performance is set at 50% and a standard of excellence is set at 85% of the total score (Alberta Education, 1999). These values translate onto the IRT theta scale score as 0.0 and 1.0, respectively. Although the Alberta Learning Achievement testing program uses cut-points, they also use a norm-referenced approach for score interpretation. Hambleton and Novick (1973) noted that, although many testing organizations have a criterion-referenced approach as they follow certain instructional blueprints used for diagnostic score interpretations, they still retain the norm-referenced approach for score interpretation. This approach is not unlike what is found in many testing situations across North America. Therefore, in the remainder of this dissertation, achievement tests will be considered criterion-referenced, although they still possess elements of a norm-referenced interpretation.

In contrast, the licensure examination considered in this study is an example of a truly criterion-based examination. These licensure examinations are from the Medical

Council of Canada and are used to certify doctors in six content areas. The cut-point for these examinations is set at a theta value of –1.3, which is the approximate passing or certification level for graduated medical doctors who are seeking entry into supervised practice. This represents the actual amount of knowledge and level of expertise that must be demanded of every medical doctor before they are allowed to practice medicine in Canada (Medical Council of Canada, 1999).

B. Item Response Theory

It has been shown that item response theory (IRT) can play a critical role in optimal item selection for criterion-referenced tests. Items, examinees, and cut-points are all put on the same scale using IRT and thus items can be optimally selected (i.e., provide maximum information) at the cut-points of interest (Hambleton & De Gruijter, 1983). In fact, with a truly criterion-referenced test, there is no need for computerized adaptive testing because only items that maximally discriminate at the cut-points of interest are required.

It is postulated in IRT that for any examinee and test item interaction there is an underlying ability or proficiency level that influences performance on that item. Examinee performance is therefore a function of both item and person characteristics. The relationship between ability and item performance can be described by a monotonically increasing nonlinear function called an item characteristic curve (ICC). As shown in Figure 2, the ICCs for both item 1 and item 2 specify that an examinee who has a high level of ability will also have a greater probability of answering an item correctly. Conversely, an examinee with a low ability has a lower probability of obtaining the correct response. These curves provide the probability of getting an item correct at each ability level across the entire ability or theta scale. Thus, with an accurate estimate of ability, subsequent examinee performance can be predicted

(Hambleton & Swaminathan, 1985). The most commonly used models are the one-, two-, and three-parameter unidimensional IRT models for dichotomously-scored items (i.e., scored 0 or 1).

Assumptions of Item Response Theory

There are three common assumptions underlying the use of one-, two-, and three-parameter unidimensional IRT models; unidimensionality, local independence, and speededness of response. The first, unidimensionality, assumes that there is only one underlying trait or ability that accounts for an examinee's responses to a test. If the data are multidimensional, then a multidimensional IRT (MIRT) model must be employed. However, since only unidimensional data and models were considered in the present research, the reader interested in MIRT models should consult Ackerman (1990, 1991, 1994), McDonald (1999), Reckase (1997), van der Linden (1996), and van der Linden & Hambleton (1997).

The second assumption underlying the use of these models is that of local independence. This assumption states that an examinee's responses to different items must be statistically independent. That is, the order of the items administered must not affect the person's performance on the test. Local independence requires that items that require a correct response to or information from a previous item or items should not be used. If the assumption of local independence is met, then the responses to each item are explained only by underlying ability. Stated mathematically, local independence is

$$\mathrm{Pr}\,ob(U_1,U_2,\ldots,U_n|\theta)=P(U_1|\theta)P(U_2|\theta)\ldots P(U_n|\theta)=\prod_{i=1}^{n}P(U_i|\theta) \quad ,$$

where θ is the ability influencing the examinee's responses to the items; U_i is the actual response of an examinee on item i = 1,2,...,n for n the number of items; and

$P(U_i|\theta)$ denotes the probability of an examinee with an ability θ having U_i=1. For example, when local independence holds, the probability of an examinee response pattern $U = (1\ 1\ 0\ 0\ 1\ 1\ 0)$, where a correct response is indicated by a 1 and an incorrect response a 0, is equal to the product of the probabilities associated with the examinee's responses to the seven individual items.

The third common assumption of the IRT models is that of speededness, which is related to unidimensionality. It is necessary that all examinees have enough time to attempt all items so that only their ability level affects their responses to each item and not the failure to reach an item. It is implicit in the unidimensional assumption that only one ability is being measured and not a second ability of speed in answering a question (Hambleton & Swaminathan, 1985).

Item Response Models

The one-parameter logistic model (or Rasch model) is an IRT model for which it is further assumed that all items have equal discriminating power and that guessing is zero. The single item parameter that is estimated is the difficulty or *b*-parameter of an item. The equation for the one-parameter logistic model (1PL) is

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

where $P_i(\theta)$ is the probability that an examinee with an ability θ will respond correctly to dichotomously-scored item *i*, b_i is the difficulty parameter for item *i*, *a* is the discrimination parameter (which is assumed to be constant across all items), and 1.7 is the scaling factor that places the outcome closer to the scale of the normal ogive model. The ICC for this model is shown in Figure 2. The curved line represents the probability of a correct response to a multiple-choice item or any item that is scored either correct or incorrect across the range of ability values. The b_i is the value of θ on the point on the ability scale at which an examinee has a 50% chance of answering the item *i* correctly. It occurs at the point of inflexion of the ICC. In Figure 2, item 1 has a *b*-parameter of 0 while item 2 is more difficult with a *b*-parameter of 1. Since item 2 is more difficult then item 1, the curve for item 2 is shifted to the right of the item 1. Item difficulty is reported on the same scale as ability, which allows for a direct comparison between a person's ability and item difficulty. Note that the theta or ability scale is a standardized scale with a mean of 0 and a SD of 1 (Hambleton & Swaminathan, 1985).

The two-parameter logistic model (2PL) is a more general model than the one-parameter model and has item characteristic curves that vary in both difficulty (b_i) and discrimination (a_i; see Figure 3). The discrimination is the slope of the item characteristic curve at the point of inflexion. The mathematical model is

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

where $P_i(\theta)$ is the probability that an examinee with ability θ will respond correctly to a dichotomously scored item *i* with a difficulty b_i, discrimination a_i, and a scaling factor of 1.7. Figure 3 shows the ICCs for item 1 and item 2 for the 2PL model. Item 1 has a discrimination index of 1 and item 2 has a discrimination index of 0.5. A steeper slope signifies that the item is more highly discriminating, thus allowing better separation of examinees with different ability levels. This difference, in turn, will allow for more precise measurement at specific ability levels. For example, item 1 is more highly discriminating (i.e., greater slope) then item 2 and therefore has more power to distinguish between a low ability examinee and a high ability examinee. Stated another way, items with high discrimination parameters are better at separating low ability groups from high ability groups on the score scale (Hambleton & Swaminathan, 1985).

The three-parameter logistic model (3PL) was introduced in order to account for the probability of a lower ability examinee obtaining the correct response by chance to a difficult item. The equation for the 3-parameter model is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}},$$

where $P_i(\theta)$ is the probability that an examinee with ability θ will respond correctly to a dichotomously scored item *i* with difficulty b_i, discrimination a_i, pseudo-guessing parameter c_i, and a scaling factor of 1.7. Note that in the 3PL model, the difficulty parameter is still at the point of inflexion, however, the probability of a correct response is no longer at 50%. The difficulty is now halfway between the lower asymptote and unity [i.e., (1 + c_i) / 2]. The ICC for the 3PL model is shown in Figure 4. The lower asymptote or pseudo-guessing parameter is the lowest possible probability of getting an item correct. For item 1 and 2, the probability of correctly responding never falls below 0.1, even for low ability examinees (Hambleton & Swaminathan, 1985; Lord, 1980).

C. Item Banking

An item bank, also known as an item pool, is a collection of items preferably in a computer-based format. All of the items must be pre-tested and screened for quality before entering the item bank. If a testing company has previous examinations with either common items or randomly equivalent examinees across forms, then it may be possible to put these tests onto a common scale using IRT procedures and use them to start building the item bank for future administrations (Kolen & Brennan, 1995). Typically, each item entered into the bank will have information such as identification number, question and distracter text, item difficulty, item discrimination, pseudo-guessing parameter (if the three-parameter IRT model is used for item calibration), correct answer, content area, cognitive level, item writer, date of creation, and item format (i.e., dichotomous or polytomous; Millman & Arter, 1984). P-values (i.e., proportion of students who obtained the correct answer) within CTT for item bank calibration can be problematic because they are dependent on the ability level of the examinees who responded to that item (for the use of CTT for item bank development see Adema & van der Linden, 1989; Armstrong, Jones, & Wang, 1994; Lin & Spray, 2000; Sanders & Verschoor, 1998). However, using IRT calibrated items allows for a common metric to be maintained which gives greater control over the tests to be produced in the future. It is this feature that reflects the full power and utility of IRT. When using IRT, all of the item characteristics are fixed and only the abilities of the examinees are estimated. The suggested number of items to be included in the item bank is about 10 items for every one item that will be used in a test. In order to reduce cheating, an item bank should have enough items to avoid reusing the same items in a short time period (Hambleton, 1986; Millman & Arter, 1984; van der Linden, 1986; Wright & Bell, 1984; see also Carbonaro, 1988; Gierl, 1998). More recently, computer algorithms have been designed to ascertain the numerical capacity or the number of items needed to withstand overexposure (Veldkamp & van der Linden, 2000). van der Linden and Luecht (1998) suggest from experience that item banks with 500 or more items will usually produce satisfactory results in many testing situations with complicated sets of constraints.

Recent research in the area of item banking has tended to focus on sorting algorithms associated with computerized test assembly. The simultaneous assembly of parallel forms with statistical, content, cognitive level, test length, and exposure controls from item banks has only recently been advanced and thus was the impetus for this study (van der Linden & Adema, 1998). As testing organizations develop their item banks, an ever-increasing demand is applied to the already resource intensive manual assembly process. It is here where computer-assisted test assembly optimization heuristics become most applicable. If an exam is similar in structure to either the achievement test or licensure exam employed as illustrative cases in this study, then items with maximum information can be gathered around the decision cut-point of interest using an automated test assembly program like CASTISEL (Luecht, 1998). Multiple constraints can then be applied to each form allowing for total control across test forms in order to increase form parallelism. However, the automation of the test assembly process using optimization heuristics is not meant to replace the test development practitioner, but only to assist in solving the very complex test assembly problem.

The exact specifications for how large or how statistically sound an item bank should be cannot be provided because they are dependent on the number of constraints and objective functions (i.e., the statistic to be maximized, such as, information) that must be met. More specifically, the reason that there are not many standards about how large item banks should be is because there are usually too many variables to consider. Nevertheless, it is known that the shape of the item bank information function depends only on one variable—the shape of the target information function. The design and choice of the target test information functions (TTIF) is especially important in order to obtain maximum information at various cut-points along the ability continuum, thereby decreasing decision error at these points. In other words, the design aspect within the CAST model is important and will allow test developers to tailor each test according to some pre-specified criteria. Thus, building item banks close to the shape of the target information function should reduce the risk of infeasibility (Timminga, 1998). Infeasibility occurs in a model when there is no possible solution that can satisfy all possible constraints put forth by the test developer. Also, it is important to note that after the tests have been assembled via test assembly computer software, test specialists can manually add or remove items and re-assemble the tests in order to re-align the tests to the blueprints. This step will help circumvent the infeasibility problems that may occur when too many constraints are applied to automated test assembly procedures.

It should be noted that the building of high quality parallel criterion-referenced forms, with the use of optimization heuristics, relies heavily on the item bank itself. If the item bank is depleted or devoid of high quality items, then no optimal test design will produce high quality results. Therefore, testing organizations must be aware that careful planning must take place when developing and maintaining item banks. When moving from the manual assembly process to an automated assembly process it is important for the test designer to take on the new role of item bank designer in order to maximize the item bank potential and minimize the risk of infeasibility.

D. Parallel Forms Construction

In addition to computer-adaptive sequential testing, the CAST model can also be used to create ATA parallel forms. The number of parallel test forms that can be created is limited only by the number and characteristics of the items in the bank. Within this CAST framework, a test developer can generate parallel forms by constraining the first stage of the CAST assembly process and by using the same statistical target for all forms while controlling for many variables including content and cognitive-level coverage, test length, and item exposure.

Different definitions of parallelism can be found in the literature, depending on whether CTT or IRT is used and which restrictions are imposed. For example, in the CTT tradition, Gulliksen (1950) stated that parallel forms occur when they have equal means, variances, and intercorrelations (i.e., a statistical criterion). The means and variances can be compared across test forms. However, because there is only one intercorrelation between two forms, for the purposes of comparison a third test must be considered. He also describes equal validities (e.g., subject matter and item format) along with equal means, variances, and reliabilities. A chi-square test is used that simultaneously tests the hypothesis of equality of means, variances, and covariances for test parallelism (Gulliksen, 1950).

Lord and Novick (1968) stated that parallel forms must produce identical true scores and need only linearly independent errors having equal variances. They noted that this is a weaker definition compared to identical true scores and identically distributed errors of earlier definitions. They also noted that when building parallel test forms it is necessary to match the type of items, the subject matter, item difficulty, and item discrimination or item-test correlation.

Feldt and Brennan (1993) discussed parallel forms in regards to reliability estimation. If one wants to find the reliability of a measure, then one could test the same examinees a week later to see if they obtained the same score (i.e., test-retest approach). Or, one could use an alternate forms approach, which is often called the parallel-forms approach. However, the authors noted that the classic definition of parallel forms may not strictly be met.

Gulliksen was the first to introduce a graphical method to help ensure statistical parallelism across forms. van der Linden and Boekkooi-Timminga (1988) created a zero-one programming approach to Gulliksen's (1950) matched random sub-tests method. With this method items are clustered into groups that are matched on the proportion passing an item and item-test correlation, which maximizes coefficient alpha as a lower bound to the classical test reliability coefficient. Gulliksen's method produces sub-tests that have the same means and equal error variances, thus meeting the requirements for parallel measurements. van der Linden and Boekkooi-Timminga's work marks some of the earliest research in using computers to build classical statistically parallel test forms.

Sanders and Verschoor (1998) demonstrated parallel forms construction using classical test theory. They used a computer algorithm that tries to maximize reliability by choosing items with high item discrimination instead of item difficulty. They differentiated between two "greedy" algorithms (i.e., they are greedy in the sense that they search for only the most discriminating items). The first, called Quick, was developed for the construction of weakly parallel forms that had identical means, variances, and reliability coefficients. The second, called Strong, was developed to create strongly parallel forms that consisted of items with identical item parameters (i.e., difficulty and point-biserial correlation).

Additional research in the area of parallel forms construction based on CTT can be found in Adema and van der Linden (1989), Armstrong, Jones, and Wang (1994), Gibson and Weiner (1998), and Lin and Spray (2000).

Within item response theory (IRT) parallel forms can be described more strictly, where each item's characteristic curve must have an equal partner on both forms (i.e., strong parallelism). Two tests are strongly (or strictly) parallel if, and only if, they have the same number of items, score categories, and each item has a matching ICC on the alternate form (Samejima, 1977). For the purpose of clarity, McDonald's (1999) terminology is adopted to describe item-parallelism. The resulting property of item-parallelism is that of equity, which states that it should not matter to an examinee from which test their score is obtained. The outcome that results from item-parallel forms is that each form will have an equivalent test characteristic curve (TCC-parallel), equal test information function (TIF-parallel), and will thus have matched true scores and error variances (McDonald, 1999).

Within an IRT framework, the construction of two tests based on the above definition of parallelism requires matching item characteristic curves. This tends to be a laborious task even for a computer because the probability of facing infeasibility increases substantially as the bank becomes depleted. Samejima's (1977) notion of "weakly parallel" test forms was quickly adopted and has proven useful in the test construction process utilizing IRT. Forms are said to be weakly parallel if they measure the same ability and have the same test information functions (i.e., TIF-parallel). Thus, in adopting this definition it is not required that tests have the same number of items (e.g., as it is in computer-adaptive testing) or score categories (i.e., binary or graded).

Lord (1977) demonstrated various practical applications of item response theory in his work. One very important application in particular, was the use of the item information function when building tests. He noted that an important feature of the item information function is that its contribution to the test is independent of other items. It was quickly learned that the task of matching TIFs by hand across different forms was far too difficult and thus researchers turned to the aid of computers. Theunissen's (1985, 1986) research marks some of the earliest work on implementing computer programming models to build tests from an item bank based on test information functions. Ackerman (1989) also presented some of the earliest research on automated assembly of weakly parallel tests using IRT information functions and added the combination of content balancing across forms.

With this notion of weak parallelism, there has recently been an increase in the development of automated test assembly algorithms and optimization heuristics for the construction of parallel forms. These recent developments can be attributed to computerized test assembly algorithms that simultaneously match a host of statistical (e.g., test information functions) and content-related specifications across test forms (Luecht, 1998b; Timminga, 1998; van der Linden, 1998; van der Linden & Adema, 1998; van der Linden & Reese, 1998; Wightman, 1998; see special issue of Applied Psychological Measurement, 1998). Adema (1992), in describing the issue of weakly parallel forms, noted that the methods used in the early 1990's could not meet all of the necessary complexities and constraints of the actual test construction process. In regards to these earlier optimization models, Wightman (1998) also noted that the linear programming models could not simultaneously build parallel test forms; each subsequent form was less than parallel. Research in the area of parallel forms construction based on IRT can be found in Ackerman (1989), Adema (1992), Armstrong, Jones, and Kunce (1998), Lin and Spray (2000), Luecht (1998b), Stocking, Swanson, and Pearlman (1993), Swanson and Stocking (1993), Theunissen (1985,1986), van der Linden (2000), van der Linden and Adema (1998), and van der Linden and Luecht (1998).

The above section reviewed not only various definitions of parallelism that can be found in the literature for CTT and IRT but also how the definitions were realized when moving into the realm of computerized assistance. Although this was not an exhaustive list, it provides an historical overview of the developments in the automation of parallel forms test construction and the reasons for its emergence. Later under a new procedures section, the introduction of yet another term will be developed which describes the parallelism within blueprint areas across parallel forms -- blueprint parallelism.

Computer-Adaptive Sequential Testing

With the CAST procedure, blocks of items called "modules" can be assembled and reviewed to ensure content coverage while meeting multiple statistical targets. These modules then become part of a computer-adaptive testing process, in which modules with different difficulty levels are chosen to be administered to examinees. Similar to the "testlet" procedure, all examinees may be administered the same starter module and then move to the second module according to the current ability estimate determined from the starter module. This process continues until the entire group of modules or "panel" is administered to a set of examinees. Figure 5 illustrates how the CAST modules can be linked in a three stage testing sequence. There are three primary pathways that examinees may take, depending on their current ability estimate. For example, if examinees do poorly on starter module A, then they would be administered module B at stage 2. If the examinee then does well on module B, they would be administered module F at stage 3. The modules from left to right cover the same content and cognitive level constraints and only differ in difficulty level. As mentioned in the introduction, Luecht (1998a) developed the CASTISEL computer program to build CAST modules. This program, assumes that data that fit the three-parameter logistic item response theory model are available for the items in the item bank. This assumption can be met for the one- and two-parameter models by setting the *a*- or *c*-parameters to a constant. For example, to go from the three-parameter to the two-parameter model, the *c*-parameter is set to 0. When moving from the two-parameter model to the one-parameter model, the *a*-parameter is set to some average across all items.

Normalized Weighted Absolute Deviation Heuristic

The CASTISEL computer program (Luecht, 1998b) includes the normalized weighted absolute deviation heuristic (NWADH) as a variation of a "greedy algorithm" that could meet very complex test assembly constraints found in large-scale testing situations. van der Linden and Adema (1998) described the negative aspect of the greedy algorithm, namely that the algorithm loads all of the best items onto the first test and, thus, each successive test will never be quite parallel. However, the NWADH circumvents this characteristic of the greedy algorithm by including a randomization factor into the item selection process. As a result, there is an equal opportunity to choose quality items (i.e., items with high *a*-parameters and desired *b*-parameters) for each form. Luecht (1998b) also noted that after a few items have been selected, the objective functions in the revised algorithm will usually diverge enough to cause the computer program to start searching for different items across forms. The NWADH also avoids the greediness of earlier automated test assembly algorithms by not choosing the items in order of their maximum information. Instead, this algorithm divides the total information desired for the test by the total test length to produce an average item information target. This process will be presented shortly in detail.

The NWADH uses a series of locally optimal searches to build one or more parallel test forms from an item bank. The normalization procedure allows the selection of numerous objective functions to be met simultaneously. The weighted aspect gives weights (or priorities) to items within content areas that do not meet the minimum constraints. This characteristic forces less discriminating items within certain content areas to be chosen first and thus allows for items within content areas that exceed the minimum to make up the difference for those items that do not exceed the minimum. The absolute deviation is the absolute difference between the target test information function and the current function. The term heuristic describes the problem-solving technique used to choose the most appropriate solution at each stage in the test assembly process.

Before going into detail on the quantitative constraints, it is important to demonstrate what a test developer would face when trying to build exams from an item bank. Figure 6 shows 159 item information functions that would need to be sorted and then combined to match a target like the one shown in Figure 7. Figure 7 shows the sum of the item information functions for an item bank and the standard error of estimate for this bank across the entire theta scale.

When a test is being assembled from an item bank, ideally test developers want to maximize information, but in reality they must control for test length. Thus, instead of creating a test composed of all items, a test must be created that satisfies two criteria: it must be of a certain length and it must satisfy a target of test information. These two constraints are dealt with first, followed by a discussion on the content constraints and how they are dealt with. *Note that in order to describe the formulas clearly, they are specified across a single ability, when in actuality they would span several points along the theta scale.*

The NWADH chooses items by dividing the target test information function (see Figure 8), *T*, by the number of items needed in the test, *n*. Thus, the first item the algorithm selects from the bank should have *T/n* (i.e., average) information (see Figure 9). Notice in Figure 9 that the target information is divided by the number of items in the test--the algorithm is searching for an item that matches this value. In other words, for an item to be chosen at this stage, *T/n* minus the information of the item under consideration should be close to zero. This raises the question, how close to zero is close enough? As it is, the difference obtained is difficult to interpret.

One solution is to index the value by test length. To do this, the algorithm takes the proportion of absolute item differences from the target to the sum of all absolute item differences from the target from the remaining items in the bank. This value will be closer to 0, and, if the item is at least as good a fit for selection as the average of the remaining items, it will be the reciprocal of the test length. As its fit becomes closer, the value approaches 0. The NWADH turns this minimization problem (i.e., the distance between the information function of an item and a portioned target test information function) into a maximization problem. That is,

$$e_i = 1 - \frac{d_i}{\sum_{i \in R_{j-1}} d_i}, i \in R_{j-1},$$

where e_i is a statistic denoting the goodness-of-fit for an item i and d_i is the index that defines how well the item information fits with our current item search value. The d_i value is then divided by the sum of all other possible item fits and is an assessment of how the fit of the item we are currently looking at compares to all other possible items. R_{j-1} contains the remaining items in the bank that are re-indexed after the selected items have been removed. Now, instead of an indeterminate range, one can say that the value will be maximized at 1. For example, a perfect item would equal the target item information function, giving a proportion of 0 when divided by the sum of all absolute item differences: 1 minus 0 is 1, a perfect score. Thus, the algorithm compares each item choice with all other possible choices against the ideal characteristics of an average item that would satisfy the test characteristics.

For each succeeding item, the algorithm, instead of looking at the original target test information function, looks at the target test information function minus the total information of the items that have already been selected. The algorithm treats this difference as the interim target function. The algorithm then divides the interim target function by the number of items that still need to be found (i.e., the total test length minus the number of items that have been already chosen). This value gives the target for what the next item should possess in regards to information. The same procedure is then followed for every other item needed in the test. Further, d_i, in the equation above, is given by

$$d_i = \left| \left(\frac{T - \sum_{k=1}^{I} u_k x_k}{n - j + 1} \right) - u_i \right| ; i \in R_{j-1},$$

where T denotes the target test information function, the sum of u_k is the total information that is already in the test, where x_k is an index of whether or not we have already chosen item k (and will be coded as either 0 or 1), u_i is the information of the item that we are currently inspecting, n is the length of the test, and j denotes the actual number of the current item under inspection compared to the total needed. For example, if we need 10 items and have already chosen 5, $j = 6$. As the d_i values decrease, the e_i values increase and this outcome represents a closer fit to the interim target for selecting item $j = 1, \ldots n$. From the previous formula, the expression

$$\frac{T - \sum_{k=1}^{I} u_k x_k}{n - j + 1},$$

represents the target value for the selection of the next item.

The content constraints are addressed in a different way. A weighting scheme is created that prioritizes item selection within content area. A weighting scheme was devised by Luecht (1998b) to ensure optimization by starting with the weakest content area (i.e., the smallest number of items, all or most of which have low discrimination) and selecting particular items first if they did not meet the minimum constraints assigned and thereby speeding up the NWADH process. The weights are adjusted after each iteration of the NWADH with the items not meeting the minimum constraints receiving more weight than the items that meet or exceed the minimum. Luecht (1998b) called this prioritized outcome the "need-to-availability ratio" (p. 230). This ratio is computed so that every test form will have the same number of items that meet both quantitative and content constraints with priority going to the categories that have the greatest need-to-availability ratio. Content areas with few items and with items that do not meet the minimum constraints are chosen first. Consequently, the content areas with a low need-to-availability ratio will be forced to make up the difference later in the iterative process. Note that one aspect of this research will involve item bank maintenance and the detection of content areas that need item development. In practice, before tests are built from item banks, the bank should first be examined and the weaker content areas developed by content experts and item writers by adding more items to these areas.

Measures of Parallelism

Existing Procedures for Parallel Forms Assessment

Test information and the standard error of estimation. The use of IRT (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980) in automated test assembly allows for a method of item and test selection and comparison of parallel forms based on item and test information functions. One of the constraints in this study is that the target test information function (TTIF) is fixed across all forms. A TTIF represents the amount of measurement precision the test developer wants to achieve across the entire theta score scale. Statistically defined, the item information function (IIF) is inversely proportional to the square of the width of the asymptotic confidence interval for θ. This relationship implies that the larger the information function, the smaller the confidence interval and the more accurate the measurement. For the logistic IRT dichotomous models, the item information function for the two-parameter model is (Lord, 1980; see relationship between ICC and item information function (IIF) in Figure 10):

$$I(\theta) = D^2 a_i^2 P_i Q_i ,$$

and for the three-parameter logistic model is,

$$I(\theta) = D^2 a_i^2 \frac{Q_i}{P_i} \left(\frac{P_i - c_i}{1 - c_i} \right)^2 ,$$

where D=1.7, a_i is the item discrimination parameter, b_i is the item difficulty parameter, and c_i is the pseudo-chance level. P_i is the probability of an examinee at a certain θ level obtaining the correct answer to item i, with Q_i being equal to $1 - P_i$. For any given ability level, the amount of information increases with larger values of a_i and decreases with larger values of c_i. That is, item discrimination reflects the amount of information an item provides assuming the pseudo-chance level is relatively small. Figure 11 illustrates graphically the relationship between the IIF and the standard error of estimation for an item.

The test information function is an extension of the item information function. The test information function is the sum of the item information functions at a given ability level:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) ,$$

where $I_i(\theta)$ is the item information function and n is the number of test items. It defines the relationship between ability and the information provided by a test. The more information each item contributes, the higher the test information function. The reciprocal of the information function at θ_i is the asymptotic sampling variance of the maximum likelihood estimate at θ_i. The standard error of estimation for a test can then be expressed as

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} ,$$

where $I(\theta)$is the test information function. The standard error of estimation can then be used to develop confidence intervals around any particular ability estimate of interest. The standard error of estimation within item response theory is comparable to classical test theory's (CTT) standard error of measurement, except within IRT the standard error is not constant along the ability continuum. Thus, the test information function and/or standard error can be used in both the construction of tests and the examination of parallelism across forms (Hambleton, Swaminathan, & Rogers, 1991).

Test characteristic curve method. One method of parallel forms comparison is in the conversion of the test characteristic curve (TCC) from the theta scale to the true (τ) or expected score scale. Under IRT, $\tau$is the sum of the item characteristic curves,

$$\tau|\theta = \sum_{j=1}^{n} P_j(\theta) ,$$

and only holds when the item response model fits the data set of interest. $P_j(\theta)$ was defined earlier by the one-, two-, or three-parameter model in the section on IRT. This relationship is nonlinear and is thus a nonlinear transformation of θ (unbounded metric) to the true score or number-correct scale (bounded metric). It should be noted that the final contribution of θ- to τ-score scale transformation is that an examinee's θ-score on form A can also be used to predict what τ-score the examinee would get on form B, even though they did not write form B (Hambleton et al., 1991; McDonald, 1999). Figure 12 shows a test characteristic curve for a 60-item test. The x-axis is the theta scale and the y-axis is the true score scale. It should be noted that the total score was not above 52 points nor below 12 score points on the true score scale for 99% of the examinees (i.e., from –3 to 3 on the theta score scale). Multiple-choice question formats with four distracters scored using the 3-parameter model with the pseudo-guessing parameter reflects this lower bound of chance probability. This θ- to τ-score transformation has a number of key features. For example, score reporting on the number-correct scale or percentage correct scale will make it easier to report and understand the ability scores and it should also expedite the cut-point decision process for criterion-referenced tests. This will allow for a direct comparison of true scores across forms to help determine whether or not the forms are parallel. The TCCs should be comparable to one another, otherwise, there is evidence of lack of parallelism between the two forms. For example, Figure 13 and 14 demonstrate the use of the test characteristic curves by calculating the average difference between two TCCs that are supposed to be parallel to one another. Figure 13 shows two uniform TCCs and Figure 14 shows two non-uniform TCCs.

There are several ways of comparing the similarity of any two curves. These can be broadly defined into two categories: those that examine the area beneath the curves and those that examine the value of the functions at specific locations. For example, some IRT differential item functioning methods have evaluated the area between item characteristic curves (ICCs) while others have used probability measures to calculate the difference between ICCs across a set of quadrature points. Some type of average is then found for these differences, such as dividing the total differences by the number of quadrature points. As the number of quadrature points increases, the accuracy of these types of solutions in estimating the true differences between the curves increases, maximizing at infinity (Camilli & Shepard, 1994).

Relative efficiency. A comparison of the information functions from two tests will also give an indication of whether or not the forms are approximately parallel. Relative efficiency makes a comparison of the information function of one test with that of a second test in terms of a common ability scale. The formula is,

$$RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)} \quad ,$$

where $RE(\theta)$denotes the relative efficiency and $I_A(\theta)$ / $I_B(\theta)$ denotes the information of test A compared to test B over a common ability(θ). In Figure 15, the relative efficiency of test A (30 items) is compared to test B (30 items). In this example, test B is functioning as if it were 20% shorter than test A at a theta of .5. This outcome means that we would need to increase test B from 30 to 36 items (i.e., assuming we are adding parallel items at a theta of .5) in order to produce the precision of the ability estimates of test A. Thus, relative efficiency will aid in the assessment of parallel forms construction by allowing for a direct comparison of each new automated form with the target test information function as well as with the other automated forms (Hambleton et al., 1991).

Mean TIF difference. The mean TIF difference offers researchers a numerical comparison for test fit. This measure allows the researcher to compare the TIF of the automated test and the target TIF. The mean TIF difference indicates the relative fit between the observed and the target TIF. The mean fit is the average of the deviations and the deviations can be calculated with the expression,

$$\frac{T(\theta_k)-\Sigma_i I(\theta_k;\xi_i)}{k},$$

where $T(\theta_k)$ is the target test information function, I is the item information function, and ξ_i is the parameter associated with item $i = 1, \ldots, 50$ averaged across $k = 1,\ldots,13$ quadrature points for the achievement tests for θ ranging from –3 to +3. For example, the licensure exam modules in this study, consist of $i = 1,\ldots,28$ items computed across $k=1,\ldots,17$ quadrature points for θ ranging from -8 to +8. However, non-uniform curves (or curves that cross over) may inadvertently cancel each other out and result in a zero or perfect fit, although they may not be. Thus, the mean square error of the TIF difference will only be used in determining the fit for this research. The mean square error of the TIF difference is an average of the squared deviations between the observed and target TIF (Luecht & Nungester, 1998). If two tests meet the target specifications, then they can be considered parallel. Notice that this is a comparison with each form to the target and not with each other (i.e., absolute difference).

The next chapter presents new procedures for parallel forms assessment. In contrast to the procedures just mentioned, these new procedures quantify the differences between these functions in order to numerically describe their exact relationship and thus aid test developers in the decision process of whether two forms are parallel.

CHAPTER III: New Procedures for Parallel Forms Assessment

The mean square error of TIF difference method is insufficient for the comparison of parallel forms since it does not compare tests to each other, only to the target. So, for example, if two tests are created to match a set target and if they both diverge from the target, then similarity to the target is no longer an adequate basis of comparing the two tests to each other. Also, the target is only defined at specific quadrature points and not as a function. Consequently, new methods were developed to allow for test functions to be compared. The accuracy of the quadrature point method is maximized when the number of quadrature points is infinite. Therefore, the most appropriate solution should involve determining the area between the curves using the analytic methods provided by integral calculus. Thus, the new methods involved calculating the area between the two test curves as a limit by allowing the number of quadrature points to go to infinity. These new procedures were used to supplement the old procedures outlined in the previous chapter. The mathematical derivations were created in Mathematica (1999).

Area Methods. The area methods start from the assumption that the difference between two TIF curves or two TCC curves is proportional to the area between them. Integral calculus can then be used to quantify the area between two curves. Within these area methods, there is an additional dichotomy: either the definite integral between two points along the x-axis of each function can be compared to the other as a scalar value or the integral of the function describing the absolute difference between the two curves can be assessed in terms of its deviation from zero. The former technique assumes that the criterion of similarity is to have similar areas beneath the curves, with negative differences having the ability to "outweigh" positive differences.

The latter technique was adopted here and assumes that the criterion for similarity is for the two curves to match at each point along the theta scale in our range of interest.

There are two curves of interest, both of which describe the properties of a test: the TCC and the TIF. Both curves illustrate identical data but the value of the TCC is more sensitive to variations in the *b*-parameters of the constituent items while the TIF is more sensitive to variations in the *a*- and *c*-parameters of the items. By evaluating differences in both the value and area of the TCC and the TIF, we will have a more accurate understanding of how any two tests differ.

For the purposes of this study, the criteria for parallelism of curves were: a) similarity of curve values at a target or criterion ability range and b) the matching of curves at all points along the ability scale where examinees are expected to be. The justification for the first criterion is that the automated test assembly algorithms aim to create tests that have equivalent information at specific ability levels. Therefore in order to assess how well this was accomplished, the value of the resultant test functions at these ability levels for differences were examined. The second criterion examines the ability of these tests to discriminate between examinees at all ability levels. Since it is often the case that criterion-referenced exams are also subject to norm-referenced interpretations, it is important to examine how two tests, which are constructed to be parallel at a specific ability level, differentially function at all ability levels.

The area methods examined in the present study involve two types of numerical comparison. The first uses a score referenced interpretation: the tests are compared and the differences between them are quantified on a known scale so that the differences may be interpreted in the context of the consequences of the differences. Although the area between two functions can be computed through integration, this area is not very informative since its mathematical definition spans an infinite number of points. However, dividing this precise area by the range of ability levels of interest yields an average difference between the two tests in terms of the test information, for TIF, or number correct scale, for TCC.

The second numerical comparison interprets the differences between the tests (TIFs and TCCs) as a ratio. The value quantifies, on average, the expected score of an examinee on one form as a ratio of their expected score another form. For example, at a specific ability level, students may score 10 points higher on one test form than on another, but if the lower score is 90, then the ratio of the two tests is .9, giving a proportional difference of .10, or 10% (true for large values of theta only). Thus, the raw differences between test forms become less important when the number of items on a test is large. For example, an average difference of five items has far greater consequences on a 10 item test than it does on a 100 item test. In a similar manner, this algorithm recognizes that differences in test information are not as severe if both tests have large quantities of information across the region of interest.

The area methods used for both the value-at-criterion method and the total-test function method involved: 1) using the three-parameter logistic item functions to produce the TIF and TCC for each test, 2) creating a function which describes either the difference or the ratio between the respective functions for comparable tests, and 3) integrating this function across a range of ability levels. The second ratio method recognizes the error of parameter estimation that is associated with IRT. For this reason, the true parameters of an item are not precisely known; therefore, the target to which the test assembly algorithm matches may not be accurately matched, were the "true" item parameters to be used. The most reasonable approach in defining how closely two tests match at a specific criterion should, therefore, allow for movement both above and below the set criterion.

The area method involved first defining the test characteristic curve (Equation 1) and information function (Equation 2). For both functions, summation of item functions is carried across all items *i*, from j_1 to j_n, where *n* is the last item in test *j*. To compare the test characteristic curves using the score scale, a function $T_{Djk}(\theta)$ was defined as the absolute difference between $T_j(\theta)$ and $T_k(\theta)$. The indefinite integral of this function was found and evaluated for the area between θ_1 and θ_2 (Equation 3). The average test score difference $E(T_{Djk})$ was found by then dividing this total area by the distance between θ_1 and θ_2 (Equation 4). The area to test score conversion was done to obtain a known metric. The same procedure was also used to compare the TIF's between tests, producing a value for $E(I_{Djk})$. For each comparison, the $E(T_{Djk})$'s and $E(I_{Djk})$'s were evaluated across the range of possible examinee abilities and plus or minus one standard deviation around the target of maximum information. *D* refers to this as a measure of the difference between test 1 *j* and test 2 *k*.

$$T_j(\theta) = \sum_{i=j_1}^{j_n} P_{ij}(\theta) \tag{1}$$

$$I_j(\theta) = \sum_{i=j_1}^{j_n} I_{ij}(\theta) \tag{2}$$

$$\int_{\theta_1}^{\theta_2} T_{Djk}(\theta) = \int_{\theta_1}^{\theta_2} \left| \sum_{i=j_1}^{j_n} P_{ij}(\theta) - \sum_{i=k_1}^{k_n} P_{ik}(\theta) \right| \tag{3}$$

$$E(T_{Djk})_{\theta_1,\theta_2} = \frac{\int_{\theta_1}^{\theta_2} T_{Djk}(\theta)}{|\theta_1-\theta_2|} \tag{4}$$

The first step in determining the non-directional ratio of the two TCC's being compared was to create a function, *f(θ)*, whose value is related to the ratio of the two functions at θ, regardless of which function assumed the greater value at a given θ (Equation 5). The inverse of this function, $[f(\theta)^{-1}]$, produces the ratio of the two functions as a value from 1 to infinity (Equation 6). By taking the reciprocal of $f(\theta)^{-1}$, we define the ratio as a value ranging from 0 to 1, equaling 1 when the functions are identical. However, since the other indices of curve similarity increase as error between the curves increases, for communication purposes, this value was reversed by taking the difference from 1 to produce a value T_{Rjk}, that increases from 0 as the curves diverged and approaches 1 when the ratio between the curves approaches infinity (Equation 7). Note that since the logistic function and its derivatives are asymptotic, this value of 1 is never reached since it requires the value of one of the functions to be 0. To calculate the average of this function across a predetermined range, a similar operation as the one outlined above was performed by integrating it between specified limits and dividing by the range of integration (Equation 8). The resulting value of $E(T_{Rjk})_{\theta1,\theta2}$ estimates the difference between 1 and the non-directional ratio of the two functions designated as R_{jk}. Again, the same sequence of operations was performed on TIF's to produce the error value for the parallelism of test information along θ ($E[I_{Rjk}]_{\theta1,\theta2}$).

$$f\left(\frac{Tj(\theta)}{Tk(\theta)}\right) = \frac{Tj(\theta)}{Tk(\theta)} + \left(\frac{T_j(\theta)}{T_k(\theta)}\right)^{-1} \qquad (5)$$

$$f^{-1}\left(\frac{Tj(\theta)}{Tk(\theta)}\right) = \frac{1}{2}\left(f\left(\frac{Tj(\theta)}{Tk(\theta)}\right)+\sqrt{-4+f\left(\frac{Tj(\theta)}{Tk(\theta)}\right)^2}\right) \qquad (6)$$

$$T_{Rjk} = 1 - \left(f^{-1}\left(\frac{Tj(\theta)}{Tk(\theta)} \right) \right)^{-1} \qquad (7)$$

$$E(T_{Rjk})_{\theta_1,\theta_2} = \frac{\int_{\theta_1}^{\theta_2} T_{Rjk}(\theta)}{|\theta_1 - \theta_2|} \qquad (8)$$

The resulting values from each comparison were arithmetically combined to produce the various area methods (ratio [Eq. 8] and difference [Eq. 4] evaluations across the entire range and around the cut-points for both TIF's and TCC's) to form a quantitative set of parameters for evaluating test parallelism.

Blueprint parallelism. When moving from handcrafted test construction (i.e., manually-made paper-and-pencil tests) to automated test assembly (ATA) procedures, testing companies want to be assured that the new test forms built via computer meet all the important characteristics of their old forms. The NWADH (Luecht, 1998b) used in this study has a built-in mechanism called a "need-to-availability-ratio" procedure that essentially weights areas with little information within the blueprints and thus these areas are chosen first in order to maximize information and allow stronger areas to make up the difference for the weaker areas. However, this algorithm does not ensure that the tests created will be blueprint parallel (i.e., each blueprint area across forms will have the same blueprint information function). What is needed for item bank maintenance is a review of the content areas for the amount of information and difficulty level within each content area. This review can be added to the design features and blueprint parallel forms can be constructed and then combined to build total test forms that would produce the same test information function across forms, thereby meeting strong parallelism more closely. This change would certainly impose more constraints on the item bank that may not be necessary for many testing programs.

Item Bank/Blueprint Mapping

Before test construction occurs from the item bank, an assessment of the bank needs to take place. It is relatively easy to review the mean item difficulty of a bank, however, to help diagnose bank problems where infeasibility may occur, construction of the blueprint information functions can be created to give a better idea of potential problem areas. However, different blueprint areas have different numbers of items and the tests produced usually require different proportions of items to be used. As an example, there are four content areas by two cognitive levels in the Alberta Learning Mathematics Grade 9 Achievement examination blueprints that will be used in this study. The eight-blueprint information functions for the achievement examination are displayed in Table 1 with the actual percentage of test-to-item bank usage. Blueprint areas 3, 4, 7, and 8 have the highest percentage of items used compared to the number of items in the item bank blueprint areas. Figure 16 was an attempt to visualize the blueprint areas using information functions. The information contained in Figure 16 represents the amount of information, location of information, the four areas with highest percentage usage rate, and the number of items in each area. The complexity of this graph does not allow for easy interpretation for item bank maintenance. Thus, new methods were explored and are presented in the Methods chapter.

Chapter IV: Method

In order to build simultaneous parallel test forms from an item bank, the normalized weighted absolute deviation heuristic (NWADH) was employed (Leucht, 1998b). The CASTISEL program (Leucht, 1998a) uses the NWADH to optimize the item assembly process and to balance the test development constraints across the forms. It is the normalization procedure that allows the heuristic to build two or more parallel test forms while meeting both content and statistical targets.

Study #1: A Criterion-Referenced Achievement Examination

Data for the first item bank came from the Grade 9 Mathematics Achievement Testing program in the Canadian province of Alberta. The Achievement Testing Program provides teachers, parents, administrators, and students with information about the student's performance in relation to the provincial curriculum standards. Like many large testing programs, the Learning Assessment Branch at Alberta Learning has a formal review to scrutinize items during the test development process. Practicing teachers are involved throughout the test development process. This process contains three general steps: item writing, field testing, and creating the final form of the test. Item writing begins when teachers are nominated to serve on the item writing committees. Alberta Learning uses a rotating selection procedure to ensure that teachers throughout the province are represented. The item writers meet several times a year to develop new items. If possible, the items are developed using a realistic context that would be familiar or topical for students in the province. The items are reviewed by content specialists (i.e., test developers and teachers) both before field testing and after.

The blueprint for these items consists of four content areas (number, patterns and relations, shape and space, and statistics and probability) and two cognitive levels (knowledge and skills; see Table 1). Two item formats are used. The first format is multiple-choice (40 items) with four response alternatives. The second format uses numerical response questions (10 items) that require the students to work through and calculate the answer to a variety of mathematical problems with no answer choices made available. The examinees are given 90 minutes to complete the examination. The actual standards are supposed to reflect the important learning objectives in Mathematics. Grade 9 Mathematics teachers from across the province are involved in the entire process and review the tests and set the standards for each examination. Alberta Learning currently creates new tests each year with a subset of common items across years for the purpose of equating. The tests are in paper and pencil format, and to date, Alberta Learning has not created an item bank using previously administered examinations. Thus, the item bank used in this study was developed specifically for this research by the author using five previously administered Grade 9 Mathematics Achievement Tests from 1995 to 1999. Each test contained 50 multiple-choice items with an item overlap of approximately 50% from one year to the next. To create an item bank for this test, the IRT parameters were estimated using BILOG 3.11 (Mislevy & Bock, 1997) with a random sample of 6000 students. IRT characteristic curve equating was conducted (Stocking & Lord, 1983) using a common-items nonequivalent groups design. All items were equated to the 1997 score scale. After equating, the item bank contained 159 items.

The criterion-referenced achievement test used in this study is characterized by an average and a high cut score—that is, the need to differentiate examinees at two points on the theta score scale. These two points are called an acceptable standard of performance and the standard of excellence. Theta scores of 0.0 and 1.0 were used for the acceptable standard and standard of excellence, respectively. Two 50-item (40 multiple-choice and 10 numeric-response items) parallel achievement test forms were assembled from the item bank because each single form of the actual Alberta achievement test contains 50 items (thus, approximately 63% of the items from the bank were used to create two parallel forms). The targets were chosen to maximize information at the acceptable standard and the standard of excellence (Alberta Learning, 1999).

In total, the achievement test had 24 constraints. Two parallel forms were created (2 constraints). Each form had to satisfy a test blueprint with equal numbers of items in four content areas across two cognitive levels (8 blueprint cells X 2 forms = 16 constraints). Test length was fixed to 50 items per form (2 constraints). Each form was created with no overlap in items (2 constraints). Each form has its own statistical target, which is the same target across tests when creating parallel forms (2 constraints).

Study #2: A Criterion-Referenced Licensure Examination

Data for the licensure examination was obtained from the Medical Council of Canada (MCC). The MCC has a Qualifying Examination Part I (MCC QEI) and a Qualifying Examination Part II (MCC QEII). The goal of these two qualifying examinations is to evaluate the competencies required for licensure of medical doctors in Canada. These competencies are deemed to be essential for the practice of medicine by all physicians in Canada. The MCC QE Part I is now administered only in computer-based format. At the time of it's conception, the exam was split into a multiple-choice Section Component (MCQ) with six disciplines of 28 questions each, for a total of 196 questions. While, there are now seven disciplines built into the design, only the first six were involved in this research. Thus, the blueprint for this test consisted of six disciplines (i.e., preventative medicine and community health, internal medicine, surgery, obstetrics and gynecology, pediatrics, and psychiatry). The multiple-choice section of the exam is arranged into a series of sections with each section balanced across discipline. The test is administered in an adaptive format with low ability examinees receiving easier items (i.e., targeted at their own ability level) and higher ability examinees receiving harder items. This adaptive structure provides each examinee with an optimal examination. Each examinee gets the opportunity to answer a small set of items (i.e., usually a set of four items) covering pre-specified content. After each section, based on how they did on the current section, they advance to a set of items that is near their ability levels. This branching technique will occur until every examinee has been administered 168 items. The maximum time allotted for this component is 3 ½ hours. The second section is the Clinical Reasoning Skills Component (CRS), which contains approximately 30 to 33 cases, each with 1-4 questions, for a total of 78-88 questions. This section of the examination is four hours in length (Medical Council of Canada, 1999). However, this research will only consider the multiple-choice questions (i.e., dichotomously scored) because the CASTISEL computer program cannot be used with polytomously-scored items at this time.

The item bank for this examination was obtained from the MCC and contains 1,973 dichotomously scored items across six disciplines. The items in the bank were calibrated with the two-parameter logistic IRT model and scaled onto the 1998 examination score scale. IRT characteristic curve equating was used with a common-items nonequivalent groups design to create the item bank (Stocking & Lord, 1983). The criterion-referenced MCC licensure examination in this study is characterized by one cut-point at the lower end of the theta score scale. The cut-point was set at a theta value of −1.3, which is the approximate passing score on this exam for recently graduated medical doctors who seek entry into a supervised practice.

The MCC computer-based testing format used was modified in the present study to fit the computer-adaptive sequential testing framework (Luecht & Nungester, 1998). Two parallel panels were implemented, as shown in the Figure 17 framework. As shown, two parallel panels were implemented. Within each panel there were six stages. There was one module at stage 1 and there were three modules at each of the remaining five stages. Each module contained 28 items, thus the examination process consisted of six stages, and involved 16 modules and 168 items (see Figure 17). Further, each module will have a parallel form attached to it that can be randomly administered to create multiple parallel panels in the computer-adaptive sequential testing process. This approach should aid in securing the items for future administrations. Figure 17 shows the 16 modules with both an A and B form for each module at every stage. Note that there are six stages and three main routes (i.e., easy, moderate, and hard) that an examinee may move through until they have received their maximum of 168 items. Each form covered the same blueprint or content specifications and met the same target information functions. Also note that it may not be feasible to have parallel panels randomly administered because of the large number of possible routes that may be encountered within this testing process. For example, using only the first panel with just 16 modules, 99 different possible routes may be encountered. It may be a better approach to administer the parallel panels at different testing administrations instead of randomly administering the parallel modules. This approach would allow content specialists to review all possible pathways an examinee may take. In fact, a balance will need to take place between the number of modules and stages in order to ensure test security and at the same time allow for test review by content specialists.

In total, the licensure exam will have 32 modules, six stages, and three main paths (i.e., easy, medium, and hard). Each of the 16 modules will have a parallel form (16 constraints) and each form has to satisfy a test blueprint with an equal number of items in six disciplines or content areas (6 disciplines X 32 forms = 192 constraints). Test length will be fixed to 168 items per form (32 constraints), each of which will be created with unique items (32 constraints). Each form will have its own statistical target, which is the same target across exams when creating parallel forms (32 constraints). Overall, the licensure exam will include 304 constraints.

Note that whether or not we administer the panels together with randomly administered modules or as separate panels on separate occasions, the construction process will remain the same. Figure 17 demonstrates the two panels, one panel in behind the other. It is not necessary that they be administered together as a set of panels, but could be administered separately, one panel each to two different groups. Also, one could reduce the number of modules. For example, 56 items could be administered in each module instead of 28 items and this could increase the likelihood of finding a feasible solution. These decisions will not be made in this research and thus it is only important to remember that many possibilities exist and the one chosen for the current study seemed to closely model the current MCC testing process.

Item Bank/Blueprint Mapping

Before test construction from the item bank is initiated, an assessment of the bank needs to take place. It is relatively easy to review the bank for difficulty and information at any particular theta level. However, to help diagnose bank problems where infeasibility may occur, construction of the blueprint information functions could be created to provide a better idea of potential problem areas. Also, by calculating the reduction in information as tests are created, developers can use these techniques to monitor the item bank quality. Thus, a graphical procedure was developed to display these numerical characteristics. This item bank blueprint mapping procedure can also be used to assess parallelism across all of the forms that are built from the item banks (i.e., test or module blueprint parallelism). Mathematica (Wolfram Research, 1999) was used for this purpose.

Parallel Forms Assessment

After the test forms from the achievement item bank and the panels from the licensure item bank were assembled, parallelism was determined. The assessment of these forms was based on the mean square error of the TIF difference, visual results of information, the newly developed TIF/TCC area quantification procedures, and blueprint parallelism. The method of relative efficiency is more appropriate for tests of different lengths that could be used to make up the difference between test efficiency and thus was not adopted in this study. Test information functions and characteristic curves will always have some degree of variability and thus values will have to be set in place for the determination of parallelism in the actual construction process. Identical TIFs or TCCs will not always result and guidelines are needed to identify meaningful differences. Even if the TIFs and TCCs did match perfectly after automated assembly, test developers and content specialist will likely make small modifications resulting in changes to the shape of the target information functions. Also, the criterion for parallelism used in this study will not necessarily hold across all testing situations (i.e., achievement and licensure examinations from different banks) and thus different criteria will need to be determined through experience in each testing system.

Chapter V: Results

This study was designed to investigate the efficiency and effectiveness of computer-assisted parallel forms construction using the CAST model for a criterion-referenced achievement test and a criterion-referenced licensure examination. The construction of the achievement examination was implemented in order to demonstrate the use and utility of the CAST model for the purpose of parallel forms construction. The construction of the licensure examination was used to demonstrate the use and utility of a CAST system for multiple parallel panels construction. The following five constraints were incorporated in the production of both the parallel forms and multiple parallel panels: content area, cognitive level, test length, item exposure, and statistical targets. In order to assess the parallelism of each form, both current and new methods for evaluating parallel forms were assessed. The assessment of the forms and panels was based on mean square error (MSE) of the TIF difference (Absolute Measure), visual results of information functions, the newly developed TIF/TCC area quantification procedures (Relative Measures), and blueprint parallelism.

Criterion-Referenced Achievement Examination

Absolute Measures of Parallelism

Table 2 contains the means, standard deviations, and MSE of the TIF differences for the two parallel forms of the achievement test constructed using the CAST model. As shown, the mean item difficulty and standard deviation of the difficulties between the two ATA forms were similar, although the standard deviation for the second form was larger indicating that the items on this form had a wider range of difficulties. The MSEs of TIF differences were small (<0.05), indicating that the observed and the target TIFs were very similar across the full ability range for both forms. However, what constitutes a good fit has not been established by the psychometric community. Using practical experience and careful consideration of both the MSE of the TIF differences and the standard error of estimation, a value of $MSE < 0.05$ was determined to represent a good fit between the target and empirical information curves.

Table 3 contains the means, standard deviations, and MSE of the TIF differences for the five forms developed and subsequently administered by Alberta Learning. These outcomes provide a basis of comparison for the ATA forms because the "handcrafted" tests were designed to be parallel across years. As shown in Table 3, the mean difficulties range from –0.246 to 0.065 while the standard deviations range from 0.773 to 0.927. While the MSE of TIF difference was small ($MSE < 0.05$), indicating good fit to the target, for the 1998 ($MSE = 0.042$) administration, the MSEs for the remaining four administrations were greater than 0.05, indicating poor fit across the full range of ability.

Figure 18 shows the target TIF and the test information functions for the two ATA forms and the five handcrafted forms created for the 1995 to 1999 administrations. It is important to note that for the five handcrafted tests, approximately 50% of the items are common from one year to the next and 25% of the items are common across a two-year period. In contrast, there were no common items in the two ATA forms constructed in the present study. When the ATA and handcrafted forms are compared, the two ATA forms had more information than three of the handcrafted tests (1995, 1998, and 1999) at the acceptable standard ($\theta = 0.0$). The ATA forms also had less information than two of the handcrafted tests (1996 and 1997) at the acceptable standard. At the standard of excellence ($\theta = 1.0$), the two ATA forms had more information than two of the handcrafted tests (1998 and 1999), equal information with one of the handcrafted tests (1995), and less information than two of the handcrafted tests (1996 and 1997).

Relative Measures of Parallelism

Although the visual results indicate where there are differences, the exact magnitude of the difference between two test forms is not known. Consequently, new numerical relative measures were developed for this study.

Table 4 contains the results using the relative measures to assess the fit between each of the Alberta "handcrafted" tests and the first ATA form. ATA 1 was chosen to be the target function because it was approximately the average function across all of the information functions. The first four columns contain information about differences in information and the second four contain the differences in true scores between pairs of tests. Within each set of four there is a ratio difference measure with information function differences expressed as a ratio, $E(I_{Rjk})_{-3,3}$ and $E(I_{Rjk})_{0,1}$ and true score differences expressed as a ratio, $E(T_{Rjk})_{-3,3}$ and $E(T_{Rjk})_{0,1}$. In contrast to the mean TIF and MSE of TIF differences, these ratios take into account how much information or how many items are in the particular tests. For example, if the two tests being compared have large amounts of information, then the differences between the information of each will have less impact because the standard error of measurement will be small. There is also a measure that directly compares the difference in information units for the information comparisons and true score differences for the test score comparisons. The information differences are represented by $E(I_{Djk})_{-3,3}$ and $E(I_{Djk})_{0,1}$; and the true score differences are represented as $E(T_{Djk})_{-3,3}$ and $E(T_{Djk})_{0,1}$. The (–3,3) and (0,1) values represent the interval along the theta continuum in which the curves were compared. The (0,1) is a comparison of the interval within one standard deviation of the cut-point. It should be noted that the interval around the two cut-scores for the Achievement exam overlapped, leading to a range from –1 to 2. This is due to the fact that the two cut-points differed by one standard deviation. Thus the interval selected was one standard deviation on either side of the two cut-points.

It was decided, based on experience in working with the ratio measures in this research, specifically with the achievement tests, that the ratio for two forms should not be above 0.05. The information and true score difference indices were judged relative to their overall information scale or their true score scale within a particular test, which is a more familiar metric and bases of comparison.

As shown in Table 4, the average TIF ratio for the interval (-3, 3) (i.e., $E[I_{Rjk}]_{-3,3}$) of the handcrafted tests information functions to the target information function varied from approximately .10 to .18. To provide a definitional framework for comparison, the proportional difference of ATA 2 to ATA 1, shown in the last row, is 0.04, which is less than the criterion value of 0.05. Thus, the handcrafted forms are proportionally far from ATA 1. Further, while the closest match between the handcrafted tests (1998 and ATA 1) had an average TIF difference ($E[I_{Dik}]_{-3,3}$) of 0.409 across the ability scale, the average TIF difference between ATA 1 and ATA2, 0.145, was 2.8 times smaller. Around the two cut-scores, the differences between the handcrafted and ATA forms were even more pronounced. While the 1998 appeared to perform adequately, with a ratio of 0.024 and a difference of 0.351, these values were still at least three times the size of the corresponding values for the two ATA forms.

As reported in the right hand side of the Table 4 and as shown in Figure 19, the handcrafted forms performed much better in the comparisons of the TCCs. Two of the average ratio values (1996 and 1997) and one average difference (1996) were lower than the ATA comparison (Table 4). In fact, the ATA 2 had a ratio value $[E(T_{Rik})_{-3,3}]$ of 0.034. However, both 1996 (0.019) and 1997 (0.029) had lower ratio values. Around the cut-point (0,1) ATA 2 did better than any of the handcrafted forms for both the ratio $[E(T_{Rik})_{0,1}]$ and the difference $[E(D_{Rik})_{0,1}]$ measures. The two best handcrafted forms were 1998 and 1999. It is likely that the improvement of the handcrafted forms on TCC comparisons is due to the dependence of the handcrafting procedure on *p*-values, which are closely related to the *b*-parameters in IRT, for item selection. Since TCC differences are most sensitive to *b*-parameter differences, attention to matching *p*-values in test construction will minimize TCC differences. Note, however, that simply by taking one test instead of another, a student could expect to achieve up to 7% lower, corresponding to a raw score difference of approximately 3 points (y-axis), on average (see Figure 19).

Blueprint parallelism. While the two ATA blueprint map differences are small, the blueprint information maps for the ATA 1 and ATA 2 forms are not identical (shown in Figure 20). The blueprint areas are listed for ATA 1 and ATA 2 by content/cognitive level in order: (a) Number Systems/Knowledge (Blueprint 1 with 17 items), (b) Number Systems/Skills (Blueprint 2 with 36 items), (c) Patterns and Relations/Knowledge (Blueprint 3 with 10 items), (d) Patterns and Relations/ Knowledge (Blueprint 4 with 35 items), (f) Shape and Space/Knowledge (Blueprint 5 with 12 items), (g) Shape and Space/Skills (Blueprint 6 with 28 items), (h) Statistics and Probability/Knowledge (Blueprint 7 with 11 items), and (i) Statistics and Probability/Skills (Blueprint 8 with 10 items). The width of each information function represents the number of items in the blueprint area (see Table 1). The x-axis is the ability scale and the y-axis is the amount of information in each of the blueprint areas. The graphical functions of Mathematica (Wolfram Research, 1999) were used to produce the curves from the data. Inspection of these blueprint information functions reveals that blueprint areas 4, 5, 7, and 8 differ between ATA 1 and ATA 2.

Figure 21 displays the blueprint maps for the 1995 through 1999 handcrafted Achievement examinations. As shown, 1995, 1996, and 1997 had the most information, while 1999 has the lowest amount of information. Across the 1995 and 1996 Achievement examinations, blueprint areas 1, 2, 4, and 6 are different. Across the 1996 and 1997 Achievement examinations, blueprint areas 1, 2, 3, 4, and 6 are different. Across the 1997 and 1998 Achievement examinations, blueprint areas 4 and 6 are different. Across the 1998 and 1999 Achievement examinations the blueprint areas 1 and 2 are different. It should be noted that each consecutive year has a 50% overlap in items. In contrast there was no overlap in items across the ATA 1 and ATA 2 Achievement examinations. Thus, using the CAST model, two achievement examinations were built (ATA 1 and ATA 2) that met , with perhaps one exception, all of the constraints imposed in this study. However, blueprint parallelism was not necessarily strictly met (did not match perfectly) and it may be important to incorporate the blueprint parallelism aspect into the computer program CASTISEL to improve this fit.

Item Bank/Blueprint Mapping

Before tests are constructed from an item bank, an assessment of item quality and blueprint coverage should be reviewed. It is relatively straightforward to review the bank for difficulty level across a particular theta range. However, to help diagnose bank problems due to infeasibility, construction of the blueprint information functions for the full bank can be used to identify potential problem areas within the bank itself. Mathematica was used for data visualization in this part of the study. The upper graph in Figure 22 represents the Achievement item bank before any test items were removed, while the lower graph represents what was left over after two ATA forms were constructed. The x-axis is the ability scale and the y-axis is the amount of information in each of the blueprint areas with the order arbitrarily set. From front to back, the blueprint areas by content/cognitive level are in the same order as the ATA 1 and ATA 2 blueprint maps. The width represents the number of items in the blueprint area. It is readily apparent from the top graph that blueprint areas 3, 7, and 8 have low information with only a small number of items in each. It is also apparent that blueprint areas 1, 2, 4, and 6 have the highest amount of information with the largest number of items in each. As revealed by the lower graph, a third ATA form could not be easily produced because three of the blueprint areas do not have enough information following construction of the two forms. Comparison of the lower graph in Figure 22 (i.e., what is left over in the item bank) with the upper graph in Figure 22 reveals that the bank cannot assemble another test due to the lack of information remaining in blueprint areas 3, 7, and 8. In order to produce another form with the same amount of information, reuse of items would be needed or item writers would need to create more items with higher amounts of information for blueprint areas 3, 7, and 8.

The empirical (Tables 2, 3, and 4) and visual results (Figures 18 and 19) strongly indicate that the two ATA forms were parallel, with each form meeting a set content area and cognitive-level coverage, test length, item exposure limit, statistical target, and number of parallel forms. In addition, except for blueprint areas 4 and 8, the blueprint maps (Figure 20) were similar. Although other blueprint areas diverged, they do not seem to pose a problem with the small amounts of information being dealt with at this blueprint level (the SD error difference was minimal). The test information functions assumed their maximum values at the acceptable standard (i.e., theta of 0). In contrast, at the standard of excellence (i.e., theta of 1), there was less information than at the acceptable standard. This is attributable to the fact that there was a greater number of items at the acceptable standard in the item bank than at the standard of excellence thereby making it difficult to maximize information at both cut scores. To overcome this limitation, more discriminating items at the standard of excellence are needed in the bank. The combination of low discriminating and average difficulty items resulted in a parallel forms solution that missed the second target.

Criterion-Referenced Licensure Examination for First MCC Panel Assembly

Two attempts were taken to find parallel panels for the licensure examination. The first attempt involved trying to maximize the total amount of information over all stages and both panels in the test development process. This first attempt is referred to as the First MCC panel assembly in this study. The Second MCC panel assembly involved reducing the information functions at the modular level. This was done to allow the computer program to choose items at each stage with less information (these are in greater abundance), thereby allowing the computer program to find items that are more similar across modules.

Absolute Measures of Parallelism

Table 5 contains the means, standard deviations, and MSE of the TIF differences for the CAST parallel forms in the licensure examination for the First MCC panel assembly. The means and standard deviations of difficulty in Table 5 are quite similar across all A and B module pairs with the exception of 16A/B. The mean difficulties for 16 A/B were –0.645 and –1.123 and the standard deviations were 2.165 and 1.887. The MSEs are less than 0.05. However, beginning with module 13 with one exception (module 14A), the TIFs are large (e.g., 0.068 to 0.237) and the corresponding MSEs exceed 0.05. Taken together, these results indicate that while there was sufficient information to construct parallel modules through the first four stages, there was insufficient information to construct parallel modules for the remaining two stages.

The information curves corresponding to the six stages are presented in Figures 23a, b, and c. The dotted lines represent the information targets and the solid lines represent the modular information functions. There is one target information function for every pair of modules built using the CAST system. At the first four assembly stages, the modular information functions closely fit with their target information function (Figure 23a and b). In contrast, the fits between modular functions and their target function was somewhat poor for the fifth and sixth stages (Figure 23c).

Relative Measures of Parallelism

The relative measures of parallelism are reported in Table 6. This table follows the same format as Table 4, with the first four columns representing differences in module information and the second four representing differences in module characteristic curves.

The NWADH produced A and B modules that matched for modules 1 A/B through modules 12 A/B. However, beginning with module 13 A/B, the values of the ratio $E(I_{Rik})_{-8,8}$, were greater than 0.05. Further, when expressed in units of information ($E[I_{Dik}]_{-8,8}$), the largest differences across the complete theta range from –8 to 8 ranged from 0.080 to 0.231 beginning with modules 13 A/B. A similar pattern of results was obtained at the cut-scores (theta = -1.3) up through module 11 A/B ($E[I_{Rik}]_{-2.3,-.3} < 0.05$). But, beginning with module 12 A/B, the value of $E(I_{Rik})_{-2.3,-.3}$ was greater than 0.05. Around the cut-score in information units ($E[I_{Dik}]_{-2.3,-.3}$), values of 0.075 to 0.267 were obtained—somewhat large if you consider the total amount of information in modules 14 A/B, 15 A/B and 16 A/B are about 3.

For the TCC differences expressed as a ratio $E(T_{Rik})_{-8,8}$, the highest was 0.087 for both 13A/B and 16A/B panels; all other values were less than 0.05. When expressed in units of true scores the differences across the complete theta range

$E(T_{Dik})_{-8,8}$ were from 0.039 to 0.787. For the TCC differences expressed as a ratio around the cut-score, the highest was 0.113 ($E[T_{Rik}]_{-2.3,-.3}$). The largest TCC differences expressed in true score units ($E[T_{Dik}]_{-2.3,-.3}$) around the cut-score was 1.491—very large considering there is only a possible score of 28 in modules 16 A/B.

Overall, the absolute measures, the relative numerical indices and the graphs in Figures 23a, b, and c suggest that there was insufficient information to construct parallel modules at all six stages.

Blueprint parallelism. Figures 24a - f contain the blueprint maps for all of the modules in the assembly process for the multiple parallel panels. Each module is presented with its alternate form (i.e., forms A and B). The blueprint areas are listed from front to back: (a) Preventative Medicine and Community Health (Blueprint 1 with 242 items), (b) Internal Medicine (Blueprint 2 with 397 items), (c) Surgery (Blueprint 3 with 359 items), (d) Obstetrics/Gynecology (Blueprint 4 with 351 items), (e) Pediatrics (Blueprint 5 with 386 items), and (f) Psychiatry (Blueprint 6 with 237 items). The width of each of the information functions represents the number of items in the blueprint area.

Notice that the MCC modular blueprint maps are almost identical across blueprint areas for the first eight modular pairs (Figures a, b, and for the first two modular pairs in c). This implies that the information at the blueprint levels are approximately the same for each modular pair. However as the assembly process continues, the blueprint maps start to shift away from one another and are not parallel (lower modular pair in Figure 24c through f). For example, the information functions in blueprint area 1 were not the same for MCC 9A and MCC 9B (see lower modules in Figure 24c). Both 15A/B and 16A/B have almost no information across all blueprint areas, and this is a direct result of item bank depletion in the earlier stages. It should be noted here that the licensure item bank has a similar amount of information within each blueprint area, a similar number of items, and similar item characteristics (i.e., low discriminating items) which probably contributed to many of the modules ending up with similar amounts of information within each blueprint area.

Item Bank/Blueprint Mapping

An assessment of the total MCC item bank was made using the blueprint mapping procedures developed in this study. This item bank consisted of 1,973 items across 6 blueprint areas. The upper graph in Figure 25 represents the amount and location of the information for each blueprint area. The cut-point was set at –1.3. Maximum information peaks for all of the blueprint areas at this cut-point. Notice that the highest and broadest amount of information is contained in blueprint area 2 (i.e., Internal Medicine with 397 items) at about 15 units of information, which is relatively low for the large number of items. The lower graph in Figure 25 represents the amount and location of information left over in the item bank after the 32 modules were removed (i.e., 896 items were removed). The decrease in information is proportional across all of the blueprint areas, because equal numbers of items were taken from each blueprint area. Note that there are enough items left over in the item bank to build another 32 modules (two panels) but not enough information, in contrast to Alberta Achievement bank, where there is enough information for some of the blueprint areas, but not enough items.

Criterion-Referenced Licensure Examination for Second MCC Panel Assembly

Absolute Measures of Parallelism

Table 7 contains the means, standard deviations, and MSE of the TIF differences for the CAST parallel forms for the Second MCC panel assembly. When reviewing the means, differences ranged from 0.005 to 0.160 in the first five stages with slightly larger differences in Stage 6 (0.250 to 0.350). The standard deviation differences ranged from 0.008 to 0.286. The MSE of TIF differences (0.000 to 0.001) were small (MSE < 0.05), indicating that the observed and the target TIFs were comparable across the full ability range for all modules. In contrast to the First MCC panel assembly, all modules were on target.

The graphical representation of the six stages are displayed in Figures 26a to 26c. The dotted lines represent the information targets and the solid lines represent the modular information functions. As for the First MCC panel assembly, there is one target for every two modules. However, in contrast to the first attempt, the test information functions fit with minor discrepancies with their target information functions for the second attempt.

Relative Measures of Parallelism

As shown in Table 8, the values of $E(I_{Rik})_{-8,8}$ were less than 0.05 for all module pairs through Stage 5. At Stage 6, the values exceeded the 0.05 level, with 14 A/B at 0.063 and 16 A/B at 0.061. When expressed in units of information ($E[I_{Dik}]_{-8,8}$), as with the ratio measure, the first 13 module pairs had values close to zero, with the largest differences ranging from 0.035 to 0.054 in Stage 6. A maximum ratio difference of 0.058 in Stage 6 (15 A/B) was found when moving to the ratio comparisons around the cut-point ($E[I_{Rik}]_{-2.3,-.3}$). In regards to the information units around the cut-point ($E[I_{Dik}]_{-2.3,-.3}$), values of 0.022 to 0.093 were obtained—small if you consider the total amount of information in Stage 6 is about 1.6.

For the TCC differences expressed as a ratio $E(T_{Rik})_{-8,8}$, all values were less than the criterion of 0.05 for the first 5 stages. At Stage 6, two of the values exceeded the 0.05 criterion with 15 A/B at 0.052 and 16 A/B at 0.081. When expressed in units of true scores, as with the ratio measures, the differences were small through the first five stages, with the largest differences across the complete theta range $E(T_{Dik})_{-8,8}$ ranging from 0.428 to 0.508 in the sixth stage. For the TCC differences expressed as a ratio around the cut-point ($E[T_{Rik}]_{-2.3,-.3}$), only 16 A/B exceeded the criterion of 0.05. The largest TCC differences expressed in true score units ($E[T_{Dik}]_{-2.3,-.3}$) around the cut-point of interest was 0.760—small considering there is a possible score of 28 in modules 15A/B.

Overall, the absolute measures, the relative numerical indices and the graphs in Figures 26 a, b, and c suggest that there was sufficient information to construct parallel modules at all six stages.

Blueprint parallelism. Figures 27 a - f demonstrate the blueprint maps for all of the modules in the assembly process for the multiple parallel panels. Notice that the MCC modular blueprint maps are almost identical across blueprint areas for the first 12 module pairs. However, modules 13 A/B through 16 A/B did not contain blueprint maps that were the same.

Parallelism means that it should not matter to the examinee which form is administered. However, MCC 15 A and B (last stage of the medium difficulty route) demonstrate that in the sixth stage, high ability examinees who are better in Preventive Medicine and Community Health (blueprint 1) will have more weight associated with a correct score on this set of items. Thus, students who do well on the blueprint areas with high information will have more weight associated with their score. For example, let us suppose that we have two examinees that are slightly above the average student and these two students perform the same across a set of tasks. Now also suppose that examinee one is better at task one and examinee two is better at task two. Now, if task one has items that are lower in discrimination when compared to task two, then examinee one will tend to obtain a higher score than examinee two. Therefore, panels can only truly be parallel and thus fair, if the test blueprint maps are the same.

Item Bank/Blueprint Mapping

The upper graph in Figure 28 is the total amount of information and location of the information for each blueprint area. The cut-point is –1.3 and is where the maximum information peaks for all the blueprint areas. The lower graph in Figure 28 represents the amount and location of information after the Second MCC panel assembly was removed (i.e., 896 items were removed). In comparison with the First MCC panel assembly, a higher level of information overall remains in the item bank, however, not enough for another 32 module assembly.

The empirical and visual results, except for the blueprint maps, indicate the CAST forms from the MCC item bank were comparable for all of the licensure exams. In Figures 26a to 26c, almost all of the modules were on target. All of the Forms in Table 8 indicate good fit between the observed and target test information functions. In regards to the blueprint maps for both the First or Second MCC panel assemblies, parallelism across all module pairs was not met. Notice that, if the items have similar amounts of information within and across the blueprints, then when the NWADH builds the test forms, each form will end up with similar blueprint maps and thus blueprint parallelism will hold. However, in order to ensure blueprint parallelism, the computer program CASTISEL would have to be modified to incorporate this balancing of information across blueprint areas.

Additional Comments on the Relative Measures

The average area methods may overemphasize differences at the extremes of ability, and the TIF and TCC comparisons respond differentially to different test characteristics. Adjustments may be made to these procedures, including weighting the difference and ratio functions by the distribution of examinees, but caution is advised in doing so. The reason is that the mathematical operations producing these values are justifiable in that they literally measure the match of one function to another. Any weighting procedure deviates from that definition and is inherently arbitrary because the consequences of test accuracy at any ability level are not necessarily proportionate to the number of examinees at that ability level. The most important consideration is that the priority of these evaluation criteria must be defined before the numerical evaluation, because the consequences of using one criterion over another may produce different conclusions. For example, if matching tests on difficulty is the most important concern, then the TIF comparisons would be given less weight. Taking the high reliability of these values combined with the fact that their differences are a result of true differences in test characteristics, the most accurate conclusions would involve using all of the criteria with the graphical evidence and making a holistic judgment.

Chapter VI: Discussion and Conclusions

This chapter begins with a summary and placement of the current study in the area of automated test assembly (ATA). Next, a review of the findings followed by various limitations found in this research are presented, then followed by a discussion of the specific problems that occurred in this research and how these problems might be avoided in the future. The various technical aspects that would be encountered in most ATA situations are reviewed in order to provide guidance in working with ATA procedures in actual testing programs. This review is followed by a section containing a discussion of the procedures for identifying and ensuring parallelism. The future research topics that should be addressed in the area of ATA are provided in the conclusion.

Summary

Research in the area of optimal test design over the last several years has tended to focus on sorting algorithms associated with computerized test assembly problems. In fact, the simultaneous assembly of parallel forms with specific content area and cognitive-level coverage, test length, item exposure limits, statistical targets, and number of parallel forms has only recently been advanced. Two real item banks were used in the present study to evaluate the CAST procedure and ATA parallel forms construction. The first was an achievement item bank developed from five Alberta Provincial achievement examinations in mathematics. The second was a licensure item bank from the Medical Council of Canada. Both a constrained version of the computerized-adaptive sequential testing (CAST) procedure and a non-constrained version were implemented with the computer program CASTISEL and assessed in the present study. The constrained version was used to create criterion-referenced ATA parallel achievement test forms. The non-constrained version was used to create multiple parallel CAST panels. These CAST panels had six stages and three main routes (i.e., easy, moderate, and hard) that examinees may move through until they have received their maximum number of items. The forms and modules were assessed for parallelism using old and new methods designed specifically for this research. The procedures used in this research to compare different test forms were:

1. Mean Square Error (MSE) of TIF Difference: The MSE of TIF difference offer researchers a numerical comparison for test fit. This measure allows the researcher to compare the TIF of the automated test and the target TIF. This index is a measure of the absolute fit between the observed and the target TIF.

2. Visual Inspection of test information functions (TIFs) and test characteristic curves (TCCs): A visual inspection was used to find out where the two test forms or modules differed along the ability scale.

3. Relative Area Measures: There are two curves of interest, both of which describe the properties of a test, the TCC and the TIF. Both curves illustrate identical data, but the value of the TCC is more sensitive to variations in the *b*-parameters of the constituent items while the TIF is more sensitive to variations in the *a*- and *c*-parameters of the items. By evaluating differences in both value and area of the TCC and the TIF, we will have a more accurate understanding of how much any two tests differ.

4. Blueprint Parallelism: When moving from handcrafted test construction (i.e., manually-made paper-and-pencil tests) to automated test assembly (ATA) procedures, testing companies want to be assured that the new test forms built via computer meet all the important characteristics of their old forms, such as blueprint parallelism. Thus, blueprint parallelism was compared across forms visually by examining the blueprint

information functions. Note that this matching of blueprint TIFs could be implemented into the computer program CASTISEL.

This research also dealt with item bank development and maintenance. In order to help diagnose bank problems where infeasibility may occur, blueprint information functions were created. Also, by calculating the reduction in information as tests are assembled, developers can monitor the item bank quality. Thus, a graphical procedure was developed to display these numerous numerical characteristics.

As testing organizations develop their item banks, an ever-increasing demand is applied to the already resource-intensive manual assembly process. It is here where computer-assisted test assembly optimization heuristics become most applicable. For example, if an examination similar in structure to the achievement tests is required, then items with maximum information could be gathered around the decision or criterion-referenced cut scores of interest using the CAST procedure. However, the automation of the test assembly process using optimization heuristics is not meant to replace the test developer but only to assist the developer by solving very complex test assembly problems.

Results

After the test forms from the achievement item bank and the panels from the licensure item bank were assembled, parallelism was determined. For the criterion-referenced Alberta Achievement examinations, the two ATA forms developed by the CASTISEL program were parallel. The handcrafted Alberta Achievement examination from1995 through 1999 varied greatly across years with none being parallel to the target. For the two Medical Council Panel assemblies, the First MCC panel assembly resulted in the first five stages all matching their respective alternate forms with stage six missing the target. Thus, the Second MCC panel assembly was implemented by lowering the amount of target information across the modules. This lowering of information function targets allowed the sixth stage forms to come closer to their targets, although they did not actually meet the ratio difference criterion for parallelism of less than 0.05. However, in consideration of the amount of information differences at the module level, the Second MCC assembly forms were considered parallel when using all the methods. In the achievement examination section, blueprint parallelism was not met. In the Second MCC panel assembly, the blueprints in the early stages were found to be parallel, but with divergence in the latter stages. Thus, blueprint parallelism was not met for either the First or Second MCC panel assembly.

It should be noted that the interpretation of these measures is very contextual and depends on the nature of the testing program and the type of decisions that will be made. For example, Alberta Learning administers both high stake examinations (i.e., Grade 12 Diploma examinations for graduation and entrance into university) and low stake examinations (i.e., Achievement examinations in grades 3, 6, and 9 to assess how the standards are being met across the province; Alberta Learning, 1999). The criteria for comparability of forms may be different for high stake examinations compared to low stake examinations.

Limitations

One limitation of this research was the use of a computer program that could not incorporate all of the various constraints that would be needed in most testing situations. Variables that would need to be added are; (a) gender or ethnicity of the people in the test items, (b) passage bound items, (c) item position, (d) item format, (e) percent-correct difficulty, (f) item-test correlation, (g) exposure rate, (h) word count, (i) item response time, (j) item set, and (k) enemy item identification. Another limitation in case of the licensure examination was the fact that only a six stage CAST model was used. In practice, many different types of CAST frameworks would need to be tested before deciding on the best model for the program under construction. This research also did not incorporate a content review by Alberta Learning or the Medical Council of Canada. A final limitation of this study was the use of real data. The assessment of the new measures of parallelism could also have been done using a simulation study as this would have allowed for verification of the new measures, i.e., that they were performing to expectations. Consequently, a bias was found regarding the ratio measure and is discussed in detail in Appendix A. Note that this does not impact any of the results regarding test or module parallelism in this study.

Implications of Automated Test Assembly

Infeasibility Problems

The first constraint that was imposed in this study was the actual size of the item banks. The exact specifications on how large or how good (i.e., item difficulty and discrimination) an item bank should be cannot be provided because it depends on the number of constraints that must be met in a particular testing situation. As the constraints accumulate within a particular testing program, the chance of infeasibility problems occurring becomes greater (Timminga & Adema, 1996). Timminga (1998) recently noted: "For test assembly problems, the feasible region or solution space consists of all tests that meet the model constraints in the particular model. The objective function determines the 'best' test in this region. If the region is empty, then the model is infeasible" (p. 280). When creating an item bank or pool, test developers must also consider the shape of the target information function for their testing programs. Creating banks with items close to the target information function should reduce the chances of infeasibility (Timminga, 1998).

The design and choice of the targets (i.e., TTIF) is especially important in order to obtain maximum information at various cut-points along the ability continuum, thereby decreasing decision error at those points. In other words, the design aspect of the CAST procedure is important because it allows test developers to tailor each test according to pre-specified criteria. This lowering of the information function should reduce the chances of infeasibility, and this was demonstrated when moving from the First to the Second MCC information targets. Note that the six stage CAST model used in this study is not the only one that could have been used. It may be wise to reduce the number of constraints by removing some of the stages and increasing the number of items per module. Also, the values used in this study will not necessarily hold across all testing situations (i.e., achievement and licensure examinations from different banks) and thus the values of the measures will need to be determined through experience in each new testing situation. Also, it is important to note that after the tests have been assembled using an ATA procedure, test developers can manually add or remove items to the tests. These suggestions should limit infeasibility problems that may occur when too many constraints are applied using automated test assembly procedures. This flexibility should also help developers create parallel criterion-referenced tests under highly constrained conditions utilzing the ATA and CAST procedures.

In this study, blueprint parallelism was also introduced. This addition further constrained the automated parallel forms and modular process as implemented by a computer program like CASTISEL. However, it was found that, if the actual blueprints cells have similar amounts of information within the item bank, and if there are numerous items with similar discrimination, blueprint parallelism will most likely result without an algorithm modification to the CASTISEL computer program.

Sequential Nature of the NWADH

As the NWADH was used to meet the increasingly stringent objectives for the achievement tests and licensure exams in this study, quality of one form was not sacrificed in order to build another form. Therefore, the sequential nature of the NWADH could be seen as quite beneficial because it allows the test developer to quickly determine how many parallel forms or panels can be assembled according to the target test information function specified. If the solution is inadequate, the test developer could pretest more items to build up the bank , reduce the target test information function, or do both. Of course, a reduction in the target information function may result in more parallel forms but it will also reduce measurement precision. In order to create high quality parallel criterion-referenced forms with the CAST procedure, a strong item bank is required. If the item bank is depleted or devoid of an adequate number of high quality items, then no matter which optimal test design is employed the tests will be of limited quality. Therefore, careful planning must take place when developing an item bank.

Identifying and Ensuring Parallelism Across Test Forms: A Three-Step Process

Traditionally, psychometricians have used statistical definitions and criteria to operationalize parallelism. Weakly parallel forms exist, for example, when the test information functions are comparable across forms (Samejima, 1977). van der Linden and Adema (1998) argued that content and statistical targets must be met to create parallel forms. Based on the results of this study, three components must be satisfied when creating parallel test forms—statistical, graphical, and substantive. The first component is statistical evidence. For example, empirical indices such as the mean square error of the TIF difference can be computed. The MSE of the TIF difference is a measure of fit between the observed and target test information functions. The MSE of the TIF difference less than or equal to 0.05 indicates good fit to the target (absolute measure).

A series of relative measures for quantifying the differences between TIFs and TCCs were developed in this research to supplement the current procedures. These measures allow one to develop a numerical standard from which to consistently judge each parallel form--not just to their target information functions. There was high agreement among these relative methods for quantifying differences, which addressed the problem of deciphering whether or not two forms are parallel. The priority of these evaluation criteria must be defined before the numerical evaluation as the consequences of using one criterion over another may produce different conclusions. Overall, these new methods worked well, because they provided numerical values that could be used to directly and more accurately help determine the parallelism of any two forms or modules. In most cases, these relative measures resulted in similar conclusions regarding the parallelism across forms.

Second, graphical methods should be used. For example, one can examine the observed and target test information functions to see if and where the functions overlap. In addition, a second component was developed in this dissertation, blueprint parallelism, and was reviewed by plotting each blueprint area across each parallel form and module. It may be wise to think about which constraints (e.g., blueprint constraints) are the most important and then design the testing system around this framework, keeping in mind that if too many constraints are applied it is more difficult to achieve parallelism. Taking into consideration the consistency of the numerical values, combined with the fact that these differences are a result of true differences in test characteristics, the most accurate conclusions would involve using both the numerical criteria and the graphical evidence. It is important to note that the graphical methods allow us to review the theta values at which differences exist between test forms.

The third component of identifying and ensuring parallelism across forms is substantive or judgmental evidence. For example, a substantive review by content specialists can be conducted. Content specialists could review the items to ensure test cohesion within forms, high quality across forms, and adequate content coverage. As previously discussed, within the CAST design outlined in this research, 99 possible routes can be found across the six stages. Thus, combining the two panels together and then randomly combining the modules from each panel will result in a testing system that will not allow for a substantive review of all possible routes an examinee may take. This multiple panel design should therefore be split into two separate panels, which will allow not only for item security (by administering the forms on different occasions) but also for the necessary content review by test specialists.

If the results from all three types of review (i.e., outcomes from empirical indices, graphical methods, and substantive reviews) converge, then researchers and practitioners will be assured that the CAST procedure can, in fact, produce truly parallel forms using a multi-faceted conception of parallelism. If the results do not converge, then research must be undertaken to further understand the complexities of the test development process and to integrate or merge these qualities into the ATA parallel and the CAST parallel panel procedures outlined in this dissertation.

Future Research

Two real item banks were used in this study to evaluate the CAST procedure and ATA parallel forms construction. Parallelism was assessed using statistical and graphical evidence but not substantive or judgmental evidence. Therefore, the author hopes to work on phase two, which involves the review process by content specialists.

The parallel forms generated for both the achievement tests and licensure exams will be reviewed by content specialists to see if the forms have integrity. In order for a content review to be conducted, a series of questions would need to be created for the content specialists to use. These questions should include, for example: (a) how many items need to be replaced? (b) why the items need to be replaced? (c) which items should not be placed in the same test? (d) what content areas are not well covered? (e) and what is missing from each content area? The idea behind such a series or check-list would be to gather data and information that could be used across the different forms to find out how many discrepancies there are and then to determine how severe they are. This review will then lead to an assessment of the possible changes that may be required in order to help circumvent some of the problem areas detected.

It is important to note that initial contact between the test developers/content specialists and analysts should greatly benefit any automated test development procedures. Sharing information at the beginning of the process could enhance effectiveness in several areas. That is, increased communication may result in increasing test validity and improving blueprint specifications. A closer working relationship may also eliminate the necessity of the aforementioned post-test assembly content review.

In future applications of the new measures developed in this study, some inconsistencies need to be addressed. These include: the susceptibility of specific value methods to error (i.e., numerous estimation procedures were used to find a final value), the overemphasis of the average area methods of differences at the extremes of ability (where decision-making may be less important), the bias in regards to the ratio measure, and the differential responsiveness of TIF and TCC comparisons to different test characteristics.

Future research could also include another measure of parallelism -- comparisons of expected score distributions. The expected score distributions can be computed, both graphically and analytically, once the item parameters and underlying ability distribution are known. This would enable practitioners to observe both the distribution and the number of subjects that would be expected to pass at the cut-points for each of the forms in question (T. Ackerman, personal communication, August 31, 2001; Lord, 1980).

Table 1

Blueprint Coding, Total number of items in each Blueprint Area, Number of items required pretest, and the Percentage of items used per test

Blueprint Cell	No. of Items in Blueprint Area	No. of Items Required per 50 item test	Percentage used per 50 item test
1 Number Systems /Knowledge	17	5	29%
2 Number Systems /Skills	36	8	22%
3 Patterns and Relations/Knowledge	**10**	**4**	**40%**
4 Patterns and Relations/Skills	**35**	**12**	**34%**
5 Shape and Space/ Knowledge	12	4	33%
6 Shape and Space/ Skills	28	9	32%
7 Statistics and Probability/Knowledge	**11**	**4**	**36%**
8 Statistics and Probability/Skills	**10**	**4**	**40%**

Table 2

Parallel Form Results for the ATA Achievement Tests Using a Bank With 159 Items to Target

Parallel Form	No. of Items	Mean Difficulty	SD Difficulty	MSE of TIF Difference
ATA 1	50	-0.108	0.479	0.010
ATA 2	50	-0.165	0.690	0.004

Table 3

Parallel Form Results for the Traditional Achievement Tests to Target

Form	No. of Items	Mean Difficulty	SD Difficulty	MSE of TIF Difference
1995	50	0.065	0.844	**0.062**
1996	50	-0.083	0.872	**0.158**
1997	50	0.006	0.773	**0.407**
1998	50	-0.249	0.927	0.042
1999	50	-0.246	0.922	**0.262**

Table 4

Relative Comparisons of Alberta Achievement Exams with Target Information Function[a]

Form	$E(I_{Bik})_{-3.3}$	$E(I_{Dik})_{-3.3}$	$E(I_{Bik})_{0.1}$	$E(I_{Dik})_{-0.1}$	$E(T_{Bik})_{-3.3}$	$E(T_{Dik})_{-3.3}$	$E(T_{Bik})_{0.1}$	$E(T_{Dik})_{0.1}$
1995	0.159	0.883	0.031	0.448	0.051	1.399	0.073	2.644
1996	0.106	0.833	0.077	1.180	0.019	0.482	0.027	0.989
1997	0.178	1.679	0.186	3.261	0.029	0.762	0.045	1.625
1998	0.097	0.409	0.024	0.351	0.079	1.414	0.018	0.625
1999	0.152	1.208	0.199	2.901	0.090	1.712	0.013	0.461
ATA 2[b]	0.040	0.145	0.007	0.093	0.034	0.600	0.006	0.227

[a] Target Information is ATA 1

[b] Included to provide a point of comparison for interpretation purposes

Table 5

Parallel Forms Results for the MCC Modules in the First MCC Panel Assembly

Stages	Parallel Module	No. of Items	Mean Difficulty	SD Difficulty	MSE of TIF Difference
1	1 A	28	-2.265	0.492	0.001
	1 B	28	-2.236	0.502	0.001
2	2 A	28	-2.986	0.363	0.000
	2 B	28	-2.983	0.513	0.000
2	3 A	28	-1.945	0.399	0.000
	3 B	28	-1.997	0.608	0.000
2	4 A	28	-0.951	0.270	0.000
	4 B	28	-0.938	0.449	0.000
3	5 A	28	-3.069	0.538	0.000
	5 B	28	-3.058	0.670	0.000
3	6 A	28	-1.970	0.784	0.000
	6 B	28	-1.922	0.910	0.001
3	7 A	28	-0.906	0.681	0.000
	7 B	28	-0.816	0.922	0.001
4	8 A	28	-3.094	0.925	0.001
	8 B	28	-3.156	0.948	0.001
4	9 A	28	-1.957	1.026	0.001
	9 B	28	-1.907	1.065	0.001
4	10 A	28	-0.687	1.222	0.006
	10 B	28	-0.649	1.197	0.006
5	11 A	28	-3.414	0.884	0.004
	11 B	28	-3.410	1.022	0.003
5	12 A	28	-2.002	1.173	0.022
	12 B	28	-2.070	1.313	0.031
5	13 A	28	-0.573	1.396	0.093
	13 B	28	-0.980	1.566	0.145
6	14 A	28	-3.704	1.051	0.035
	14 B	28	-3.476	1.562	0.061
6	15 A	28	-2.535	1.586	0.187
	15 B	28	-2.499	1.714	0.314
6	16 A	28	-0.645	2.165	0.329
	16 B	28	-1.123	1.887	0.374

Table 6

Results from Relative Comparisons for First MCC Panel Assembly

Stages	j,k	$E(I_{Rik})_{-8.8}$	$E(I_{Dik})_{-8.8}$	$E(I_{Rik})_{-2.3.-.3}$	$E(I_{Dik})_{-2.3.-.3}$	$E(T_{Rik})_{-8.8}$	$E(T_{Dik})_{-8.8}$	$E(T_{Rik})_{-2.3.-.3}$	$E(T_{Dik})_{-2.3.-.31}$
1	1 A,B	0.021	0.017	0.008	0.019	0.005	0.053	0.007	0.125
2	2 A, B	0.021	0.008	0.007	0.014	0.004	0.039	0.002	0.047
	3 A, B	0.012	0.007	0.002	0.004	0.022	0.086	0.008	0.126
	4 A, B	0.016	0.008	0.004	0.011	0.015	0.067	0.002	0.024
3	5 A, B	0.019	0.005	0.005	0.010	0.005	0.046	0.005	0.097
	6 A, B	0.033	0.018	0.010	0.024	0.006	0.085	0.009	0.148
	7 A, B	0.036	0.016	0.002	0.006	0.012	0.154	0.017	0.217
4	8 A, B	0.047	0.028	0.051	0.109	0.012	0.093	0.005	0.110
	9 A, B	0.024	0.015	0.009	0.023	0.008	0.077	0.006	0.097
	10 A, B	0.032	0.026	0.029	0.075	0.014	0.120	0.005	0.053
5	11 A, B	0.047	0.017	0.013	0.030	0.010	0.102	0.009	0.185
	12 A, B	0.033	0.038	0.062	0.141	0.020	0.113	0.007	0.111
	13 A, B	0.150	0.142	0.070	0.132	0.087	0.682	0.110	1.492
6	14 A, B	0.161	0.080	0.142	0.233	0.021	0.405	0.036	0.793
	15 A, B	0.166	0.231	0.182	0.267	0.049	0.498	0.047	0.854
	16 A, B	0.152	0.108	0.071	0.075	0.087	0.787	0.113	1.491

Table 7

Parallel Forms Results for the MCC Modules in the Second MCC Panel Assembly

Stages	Parallel Module	No. of Items	Mean Difficulty	SD Difficulty	MSE of TIF Difference
1	1 A	28	-1.943	0.443	0.003
	1 B	28	-1.960	0.537	0.003
2	2 A	28	-2.948	0.343	0.000
	2 B	28	-2.983	0.526	0.000
2	3 A	28	-1.942	0.421	0.000
	3 B	28	-1.874	0.746	0.000
2	4 A	28	-0.969	0.516	0.000
	4 B	28	-1.048	0.593	0.000
3	5 A	28	-2.928	0.680	0.000
	5 B	28	-3.083	0.880	0.000
3	6 A	28	-1.975	0.832	0.000
	6 B	28	-1.969	0.957	0.000
3	7 A	28	-0.787	0.974	0.000
	7 B	28	-0.831	0.894	0.000
4	8 A	28	-3.126	1.107	0.000
	8 B	28	-3.136	1.382	0.000
4	9 A	28	-1.961	1.211	0.000
	9 B	28	-2.023	1.497	0.000
4	10 A	28	-0.870	1.363	0.000
	10 B	28	-0.875	1.371	0.000
5	11 A	28	-3.239	1.457	0.000
	11 B	28	-3.216	1.329	0.000
5	12 A	28	-1.887	1.455	0.000
	12 B	28	-2.101	1.845	0.000
5	13 A	28	-0.605	1.734	0.001
	13 B	28	-0.670	1.597	0.001
6	14 A	28	-3.404	1.814	0.001
	14 B	28	-3.154	1.908	0.001
6	15 A	28	-2.176	1.649	0.001
	15 B	28	-1.834	1.712	0.001
6	16 A	28	-0.327	1.844	0.001
	16 B	28	-0.681	1.893	0.001

Table 8

Results from Relative Comparisons for the Second MCC Panel Assembly

Stages	j,k	$E(I_{Rik})_{.8.8}$	$E(I_{Dik})_{.8.8}$	$E(I_{Rik})_{-2.3,-.3}$	$E(I_{Dik})_{-2.3,-.3}$	$E(T_{Rik})_{.8.8}$	$E(T_{Dik})_{.8.8}$	$E(T_{Rik})_{-2.3,-.3}$	$E(T_{Dik})_{-2.3,-.31}$
1	1 A,B	0.004	0.002	0.001	0.001	0.005	0.027	0.003	0.044
	2 A, B	0.010	0.007	0.002	0.003	0.006	0.054	0.005	0.090
2	3 A, B	0.023	0.013	0.014	0.021	0.010	0.110	0.013	0.209
	4 A, B	0.019	0.014	0.016	0.025	0.029	0.118	0.011	0.140
	5 A, B	0.036	0.019	0.016	0.022	0.024	0.238	0.019	0.370
3	6 A, B	0.016	0.013	0.011	0.016	0.006	0.044	0.006	0.096
	7 A, B	0.019	0.016	0.008	0.012	0.016	0.095	0.016	0.198
	8 A, B	0.038	0.018	0.021	0.029	0.012	0.117	0.005	0.097
4	9 A, B	0.023	0.017	0.015	0.023	0.009	0.082	0.010	0.158
	10 A, B	0.024	0.018	0.025	0.039	0.005	0.030	0.004	0.047
	11 A, B	0.028	0.008	0.005	0.006	0.003	0.035	0.003	0.052
5	12 A, B	0.024	0.016	0.007	0.011	0.037	0.296	0.026	0.409
	13 A, B	0.024	0.021	0.025	0.040	0.028	0.121	0.011	0.128
	14 A, B	0.063	0.035	0.026	0.035	0.025	0.428	0.032	0.622
6	15 A, B	0.103	0.054	0.058	0.093	0.052	0.508	0.046	0.760
	16 A, B	0.061	0.036	0.014	0.022	0.081	0.504	0.054	0.658

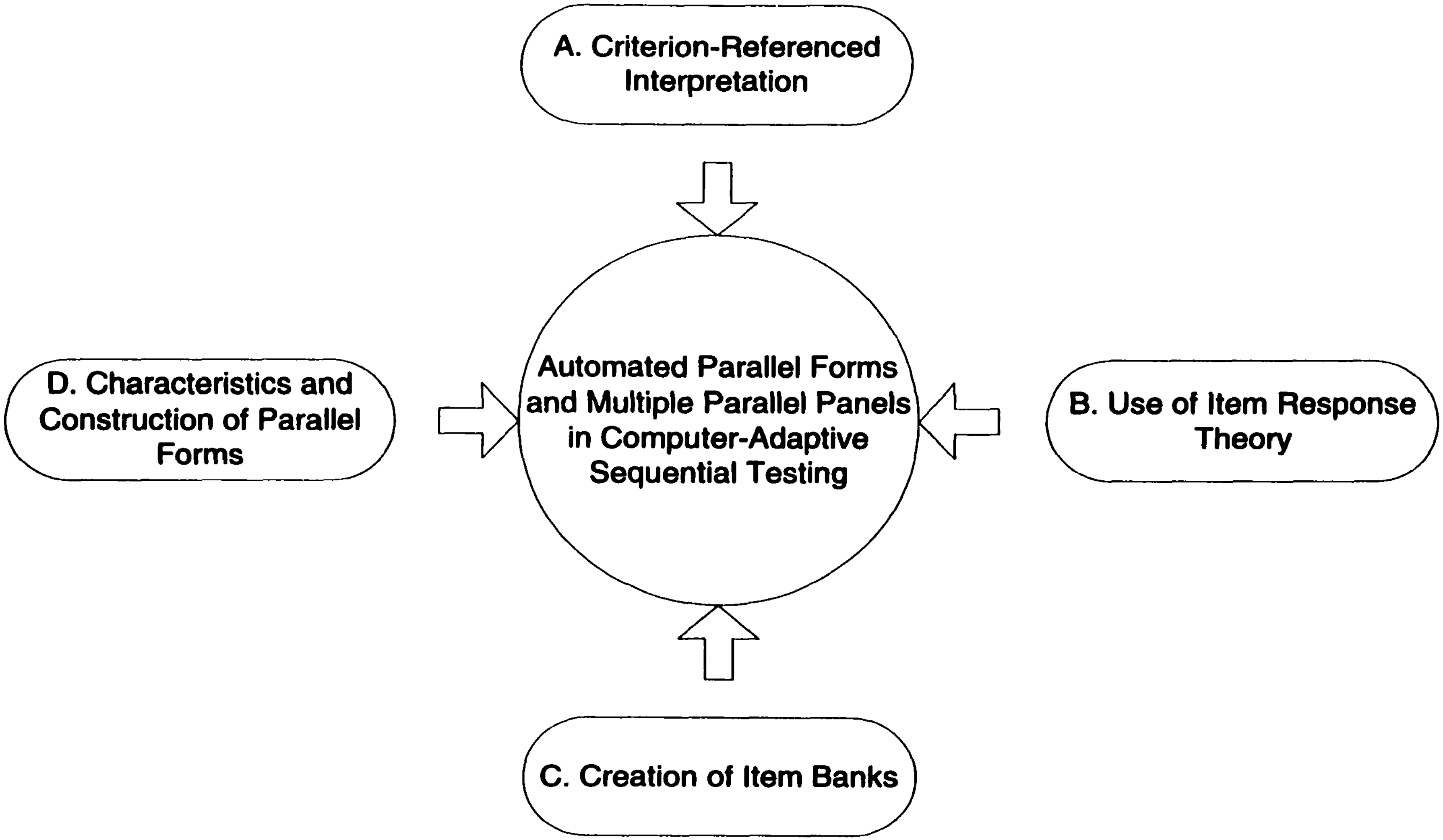

A. Criterion-Referenced Interpretation
D. Characteristics and Construction of Parallel Forms
Automated Parallel Forms and Multiple Parallel Panels in Computer-Adaptive Sequential Testing
B. Use of Item Response Theory
C. Creation of Item Banks

Figure 1. Necessary Considerations for the Construction of Automated Parallel Forms and Panels in a CAST System.

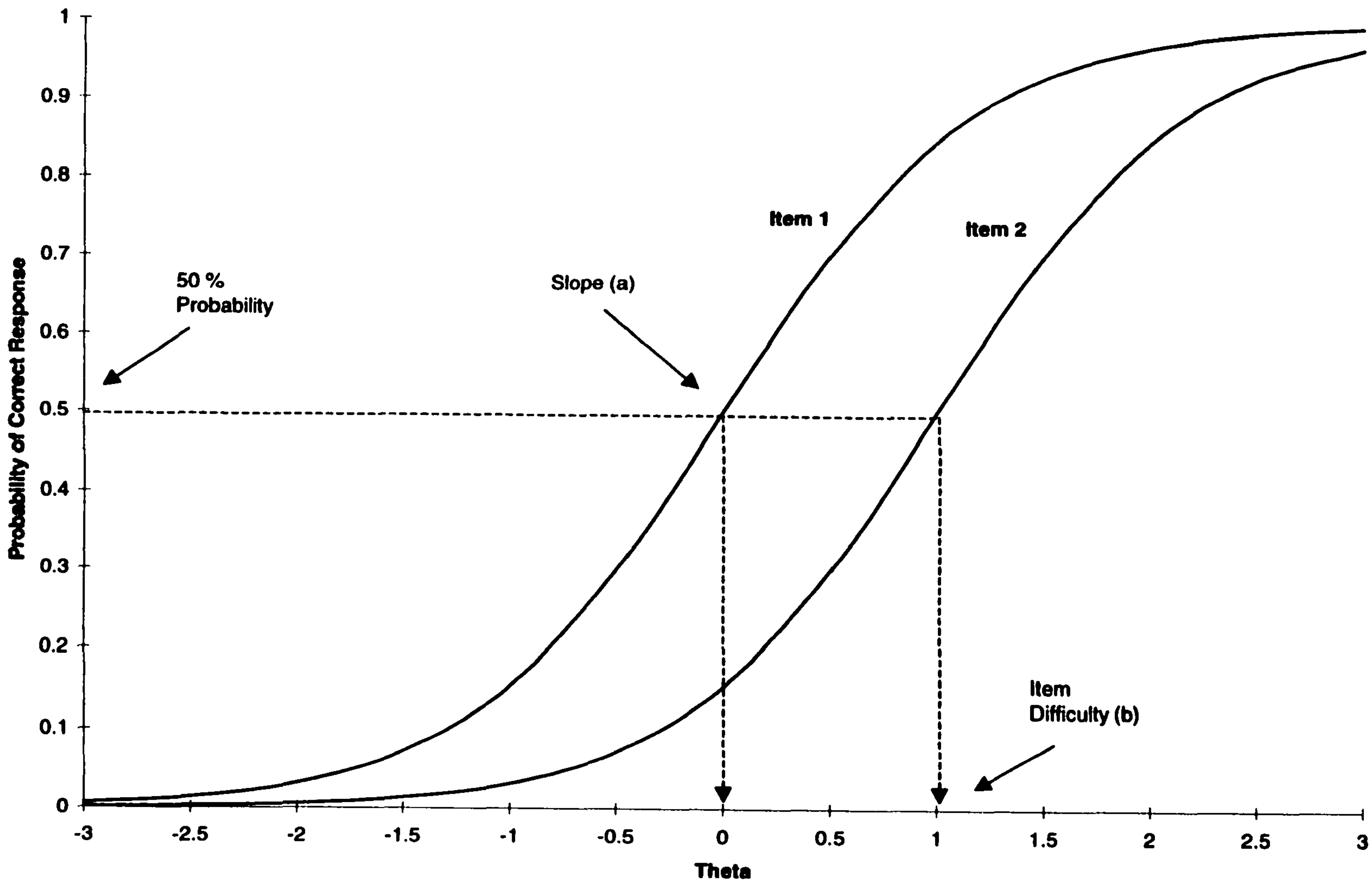


Figure 2. The One-Parameter Logistic IRT Model for Two Items.

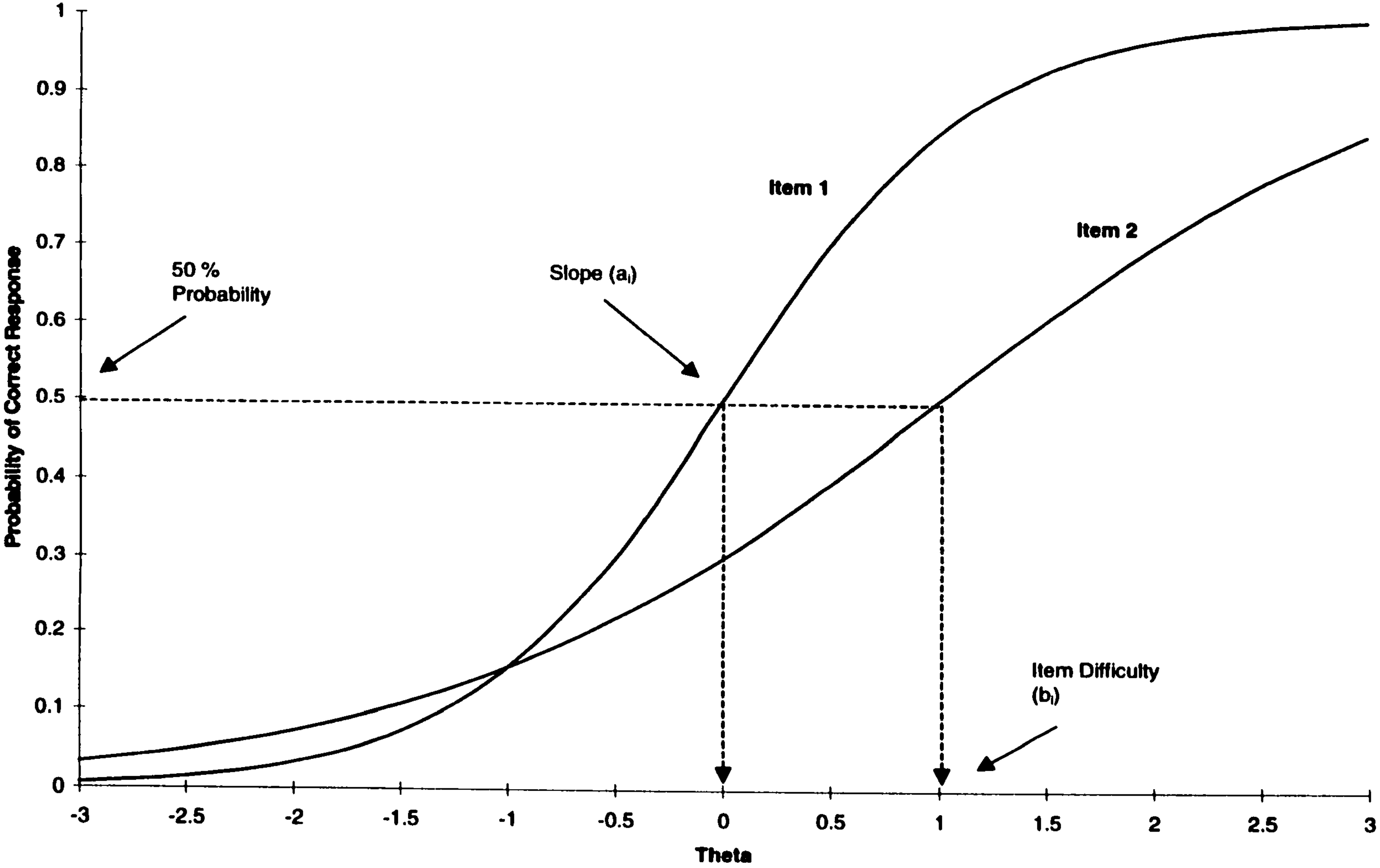


Figure 3. The Two-Parameter Logistic IRT Model for Two Items.

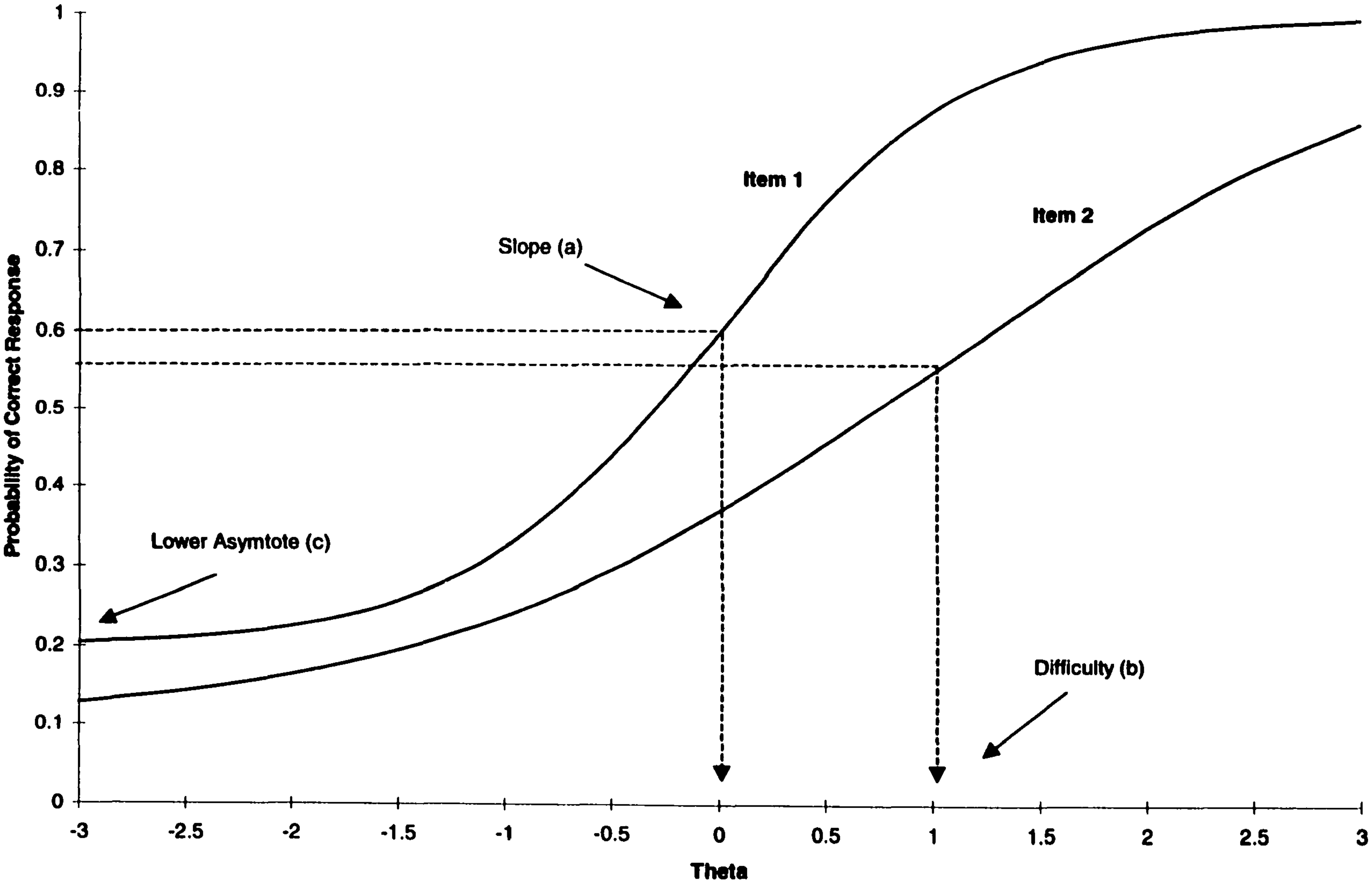


Figure 4. The Three-Parameter Logistic IRT Model for Two Items.

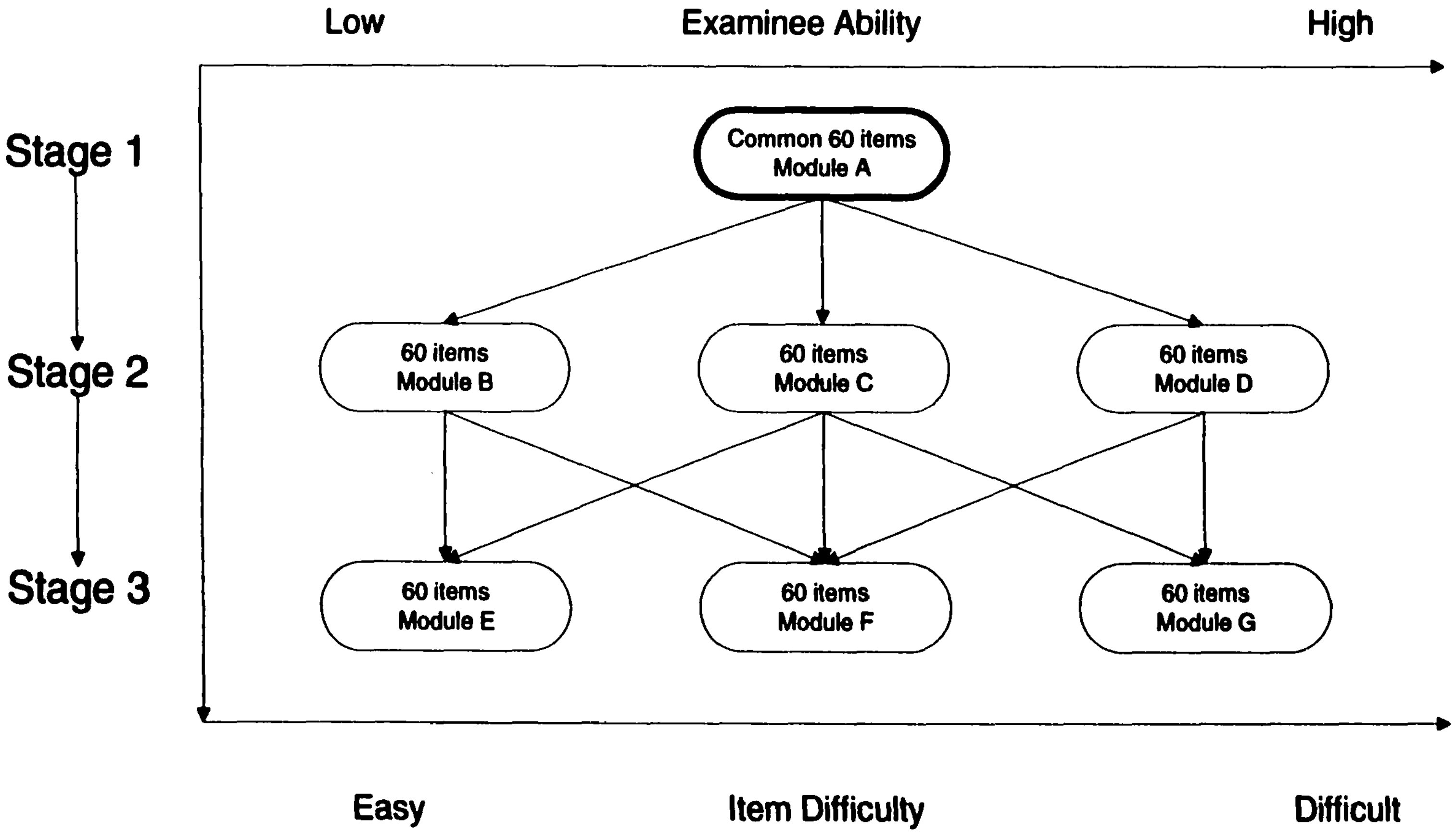


Figure 5. A Three Stage Computer-adaptive Test Example.

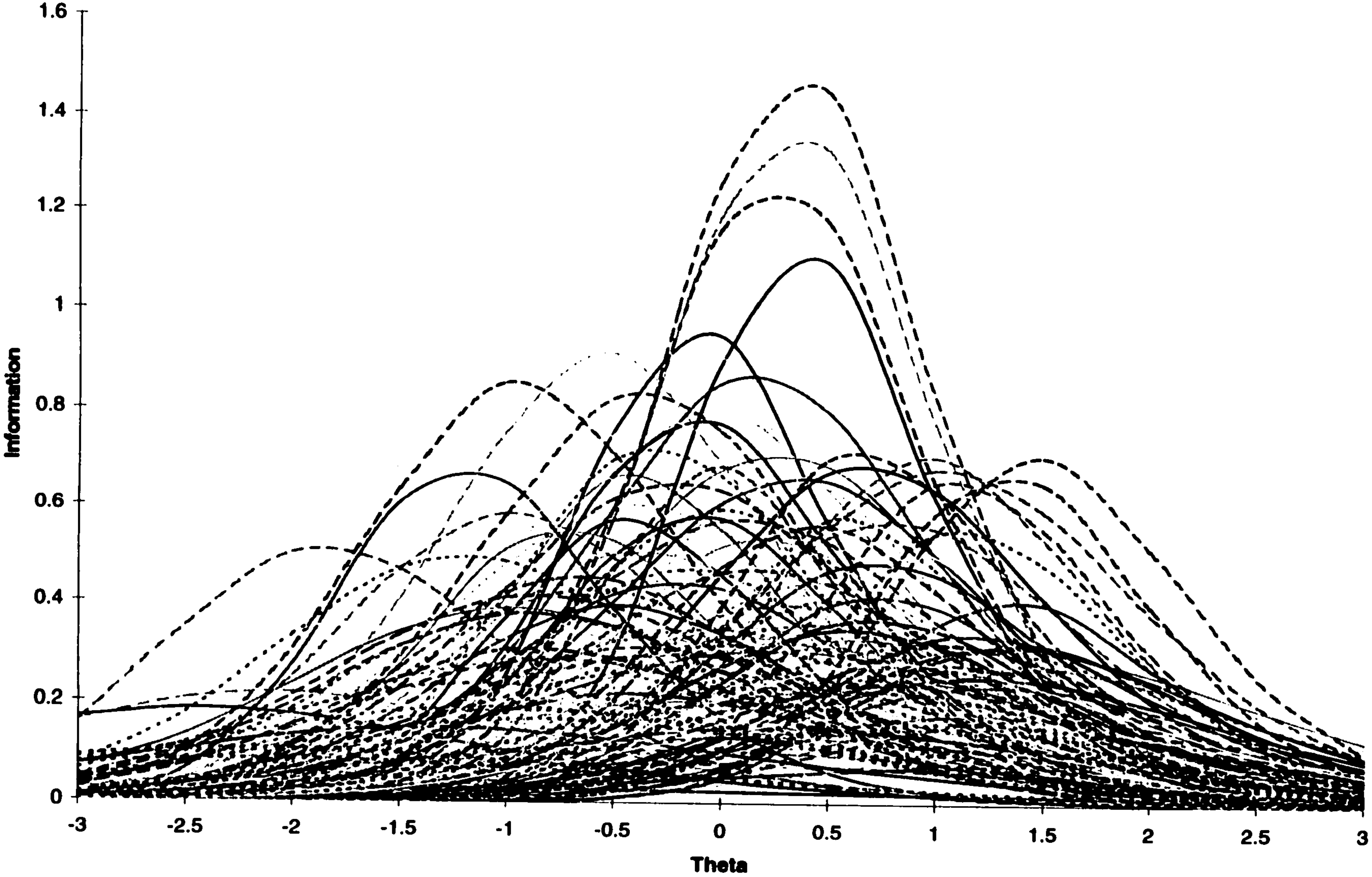


Figure 6. Item Information Functions for the Alberta Achievement Item Bank.

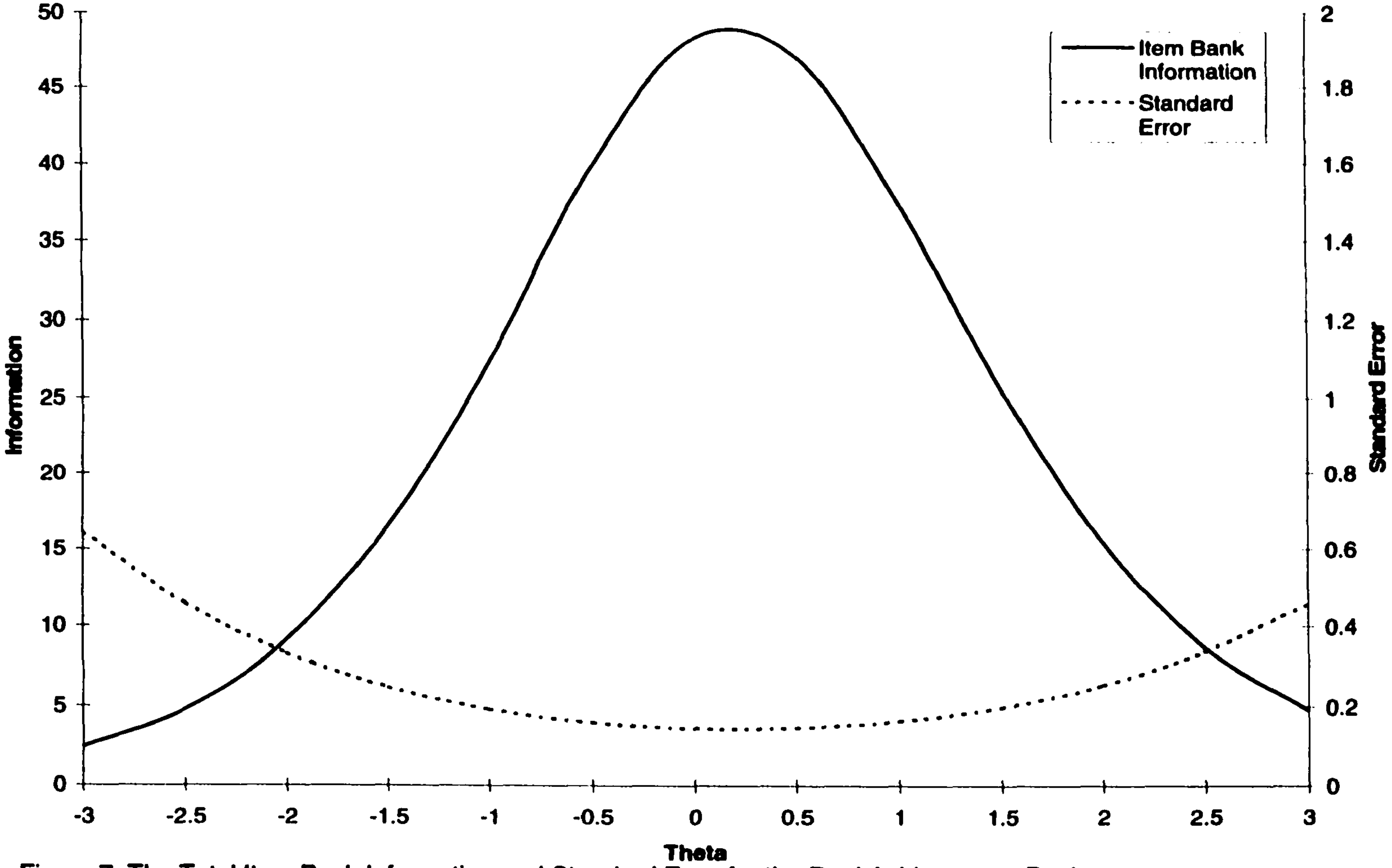


Figure 7. The Total Item Bank Information and Standard Error for the Real Achievement Bank.

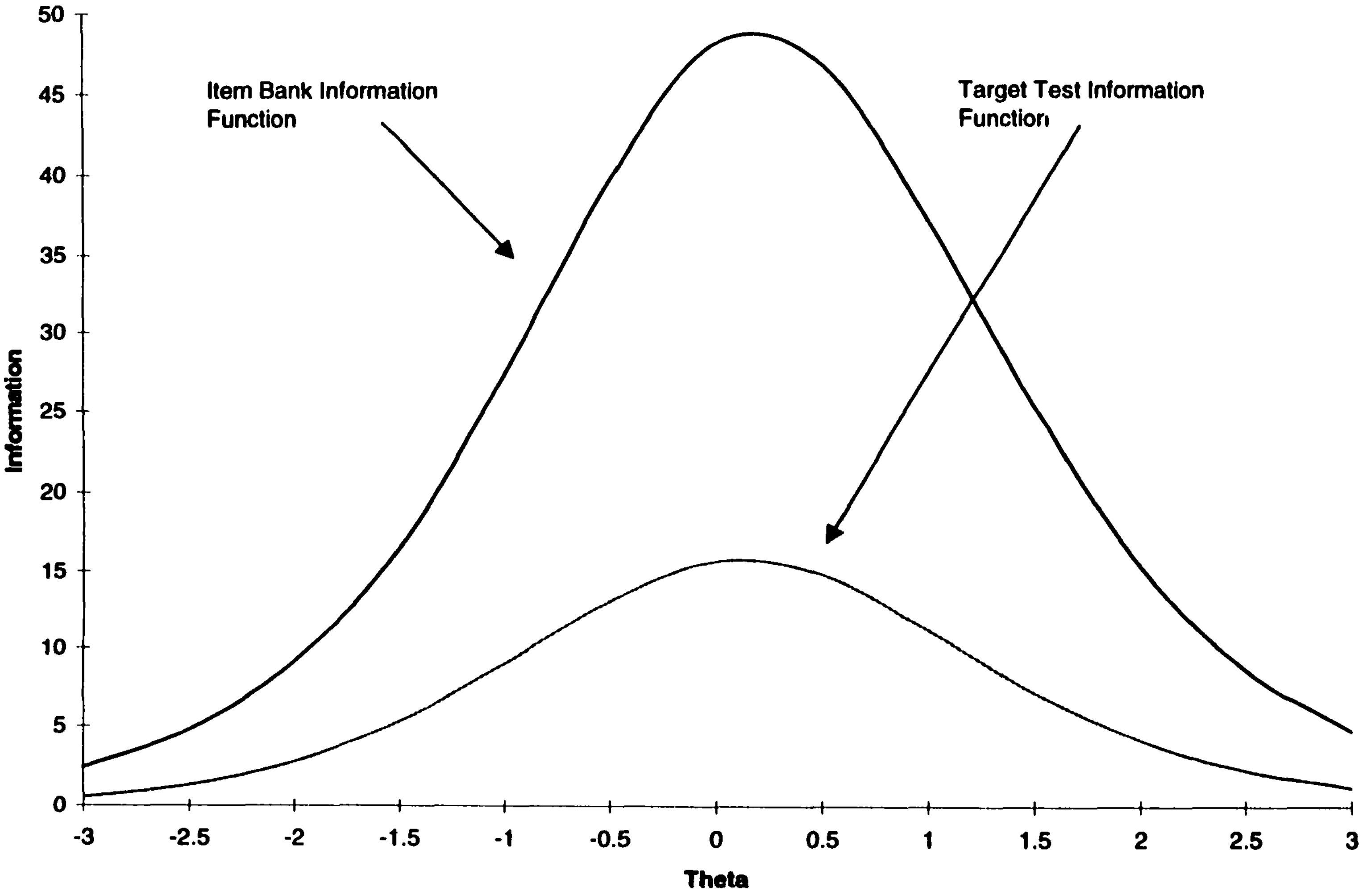


Figure 8. The Total Item Bank Information with a Target Test Information Function.

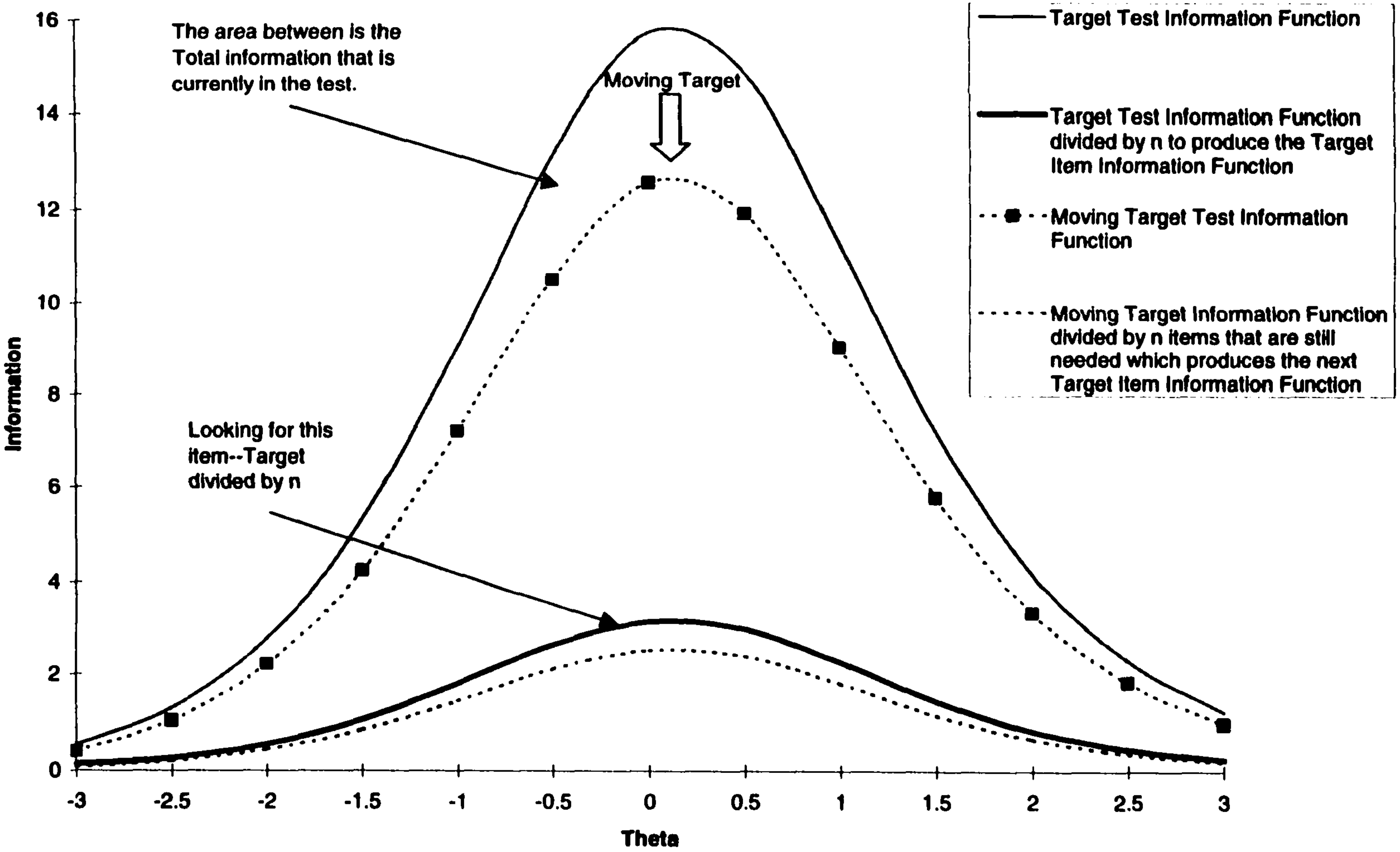


Figure 9. The Moving Target Test Information Function

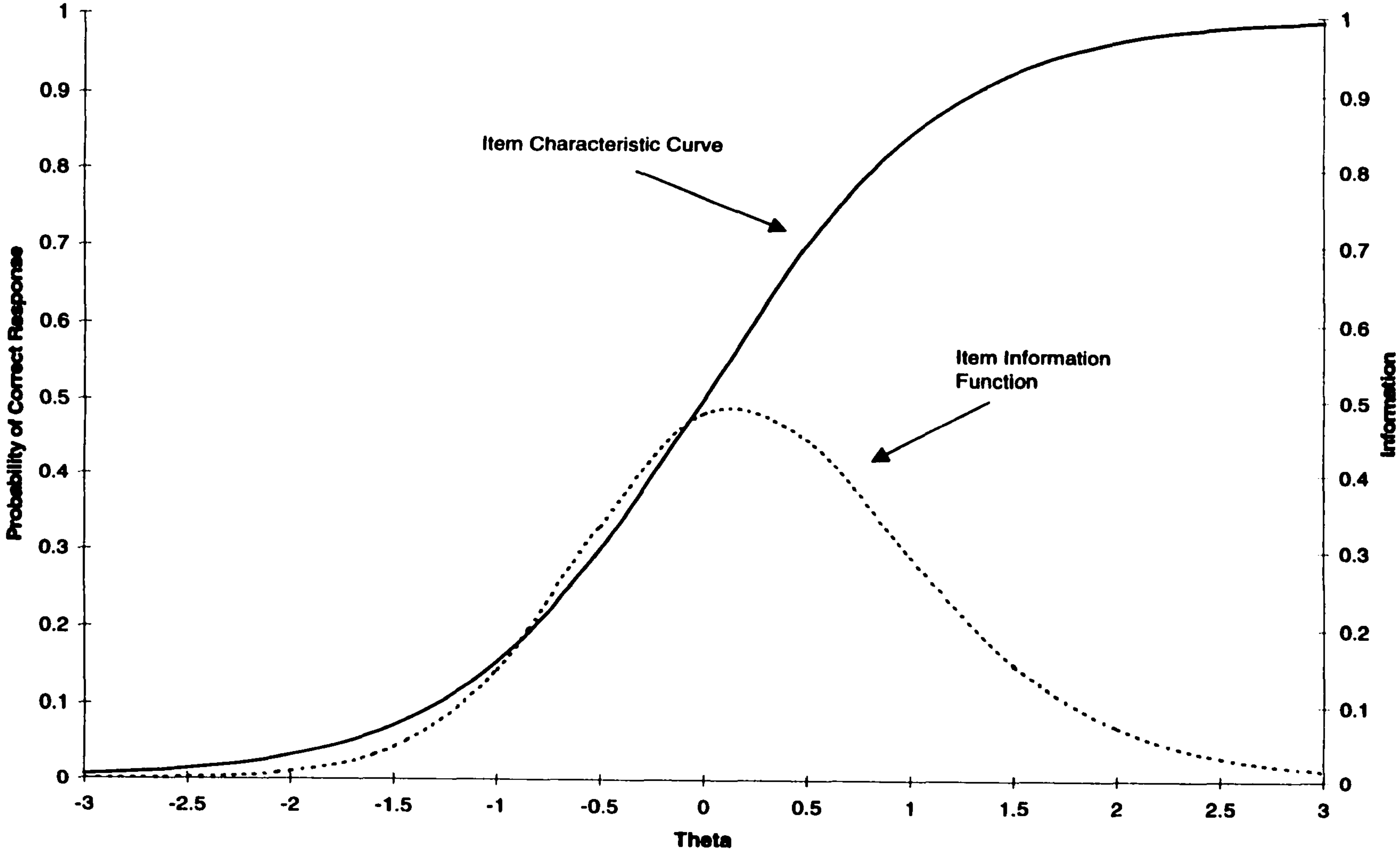


Figure 10. A Two-Parameter Item Characteristic Curve with its Information Function.

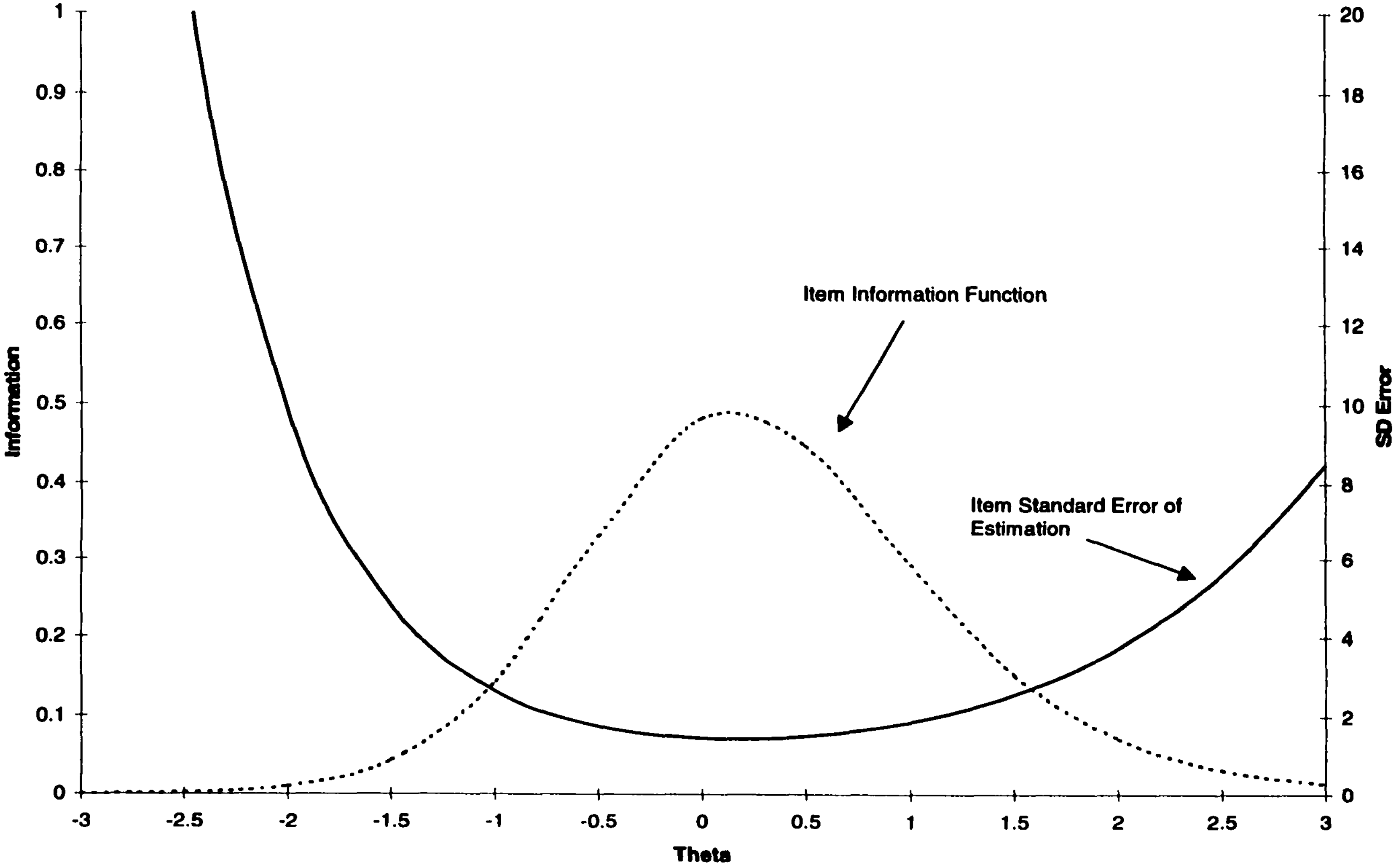


Figure 11. An Information Function with its Standard Error of Measurement.

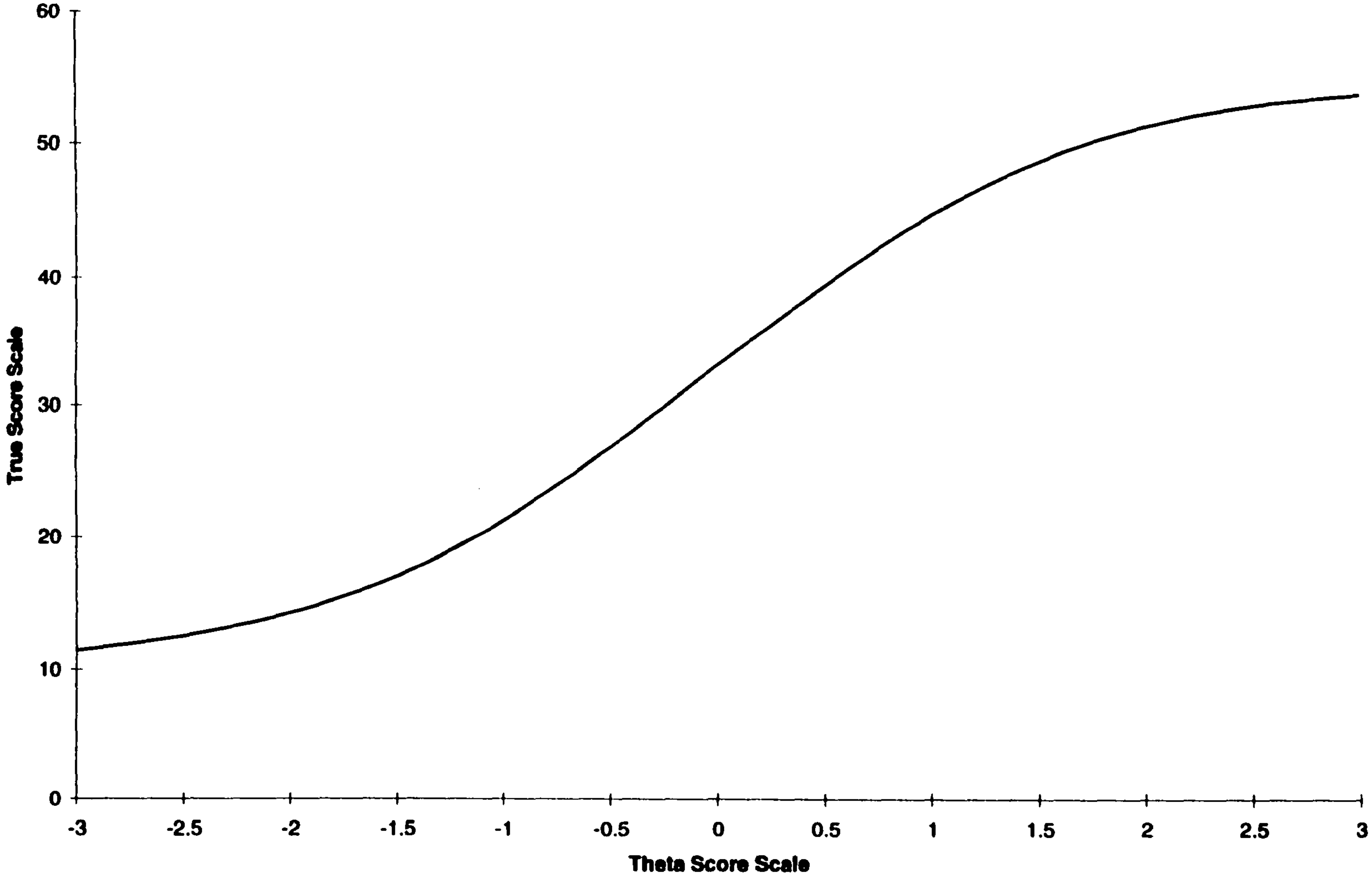


Figure 12. The Test Characteristic Curve for a 60 Item Test.

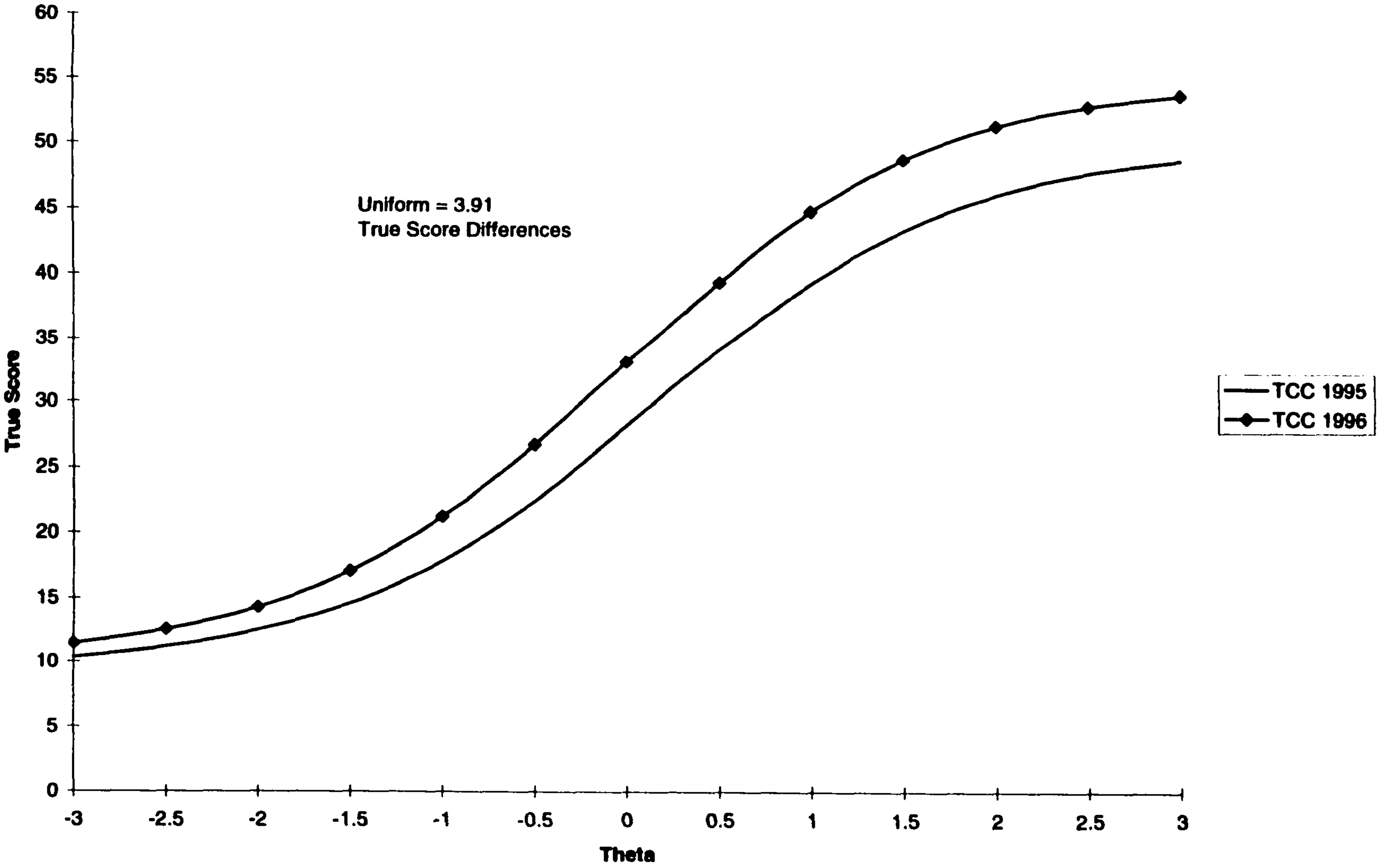


Figure 13. Two Test Characteristic Curves with the Average Difference.

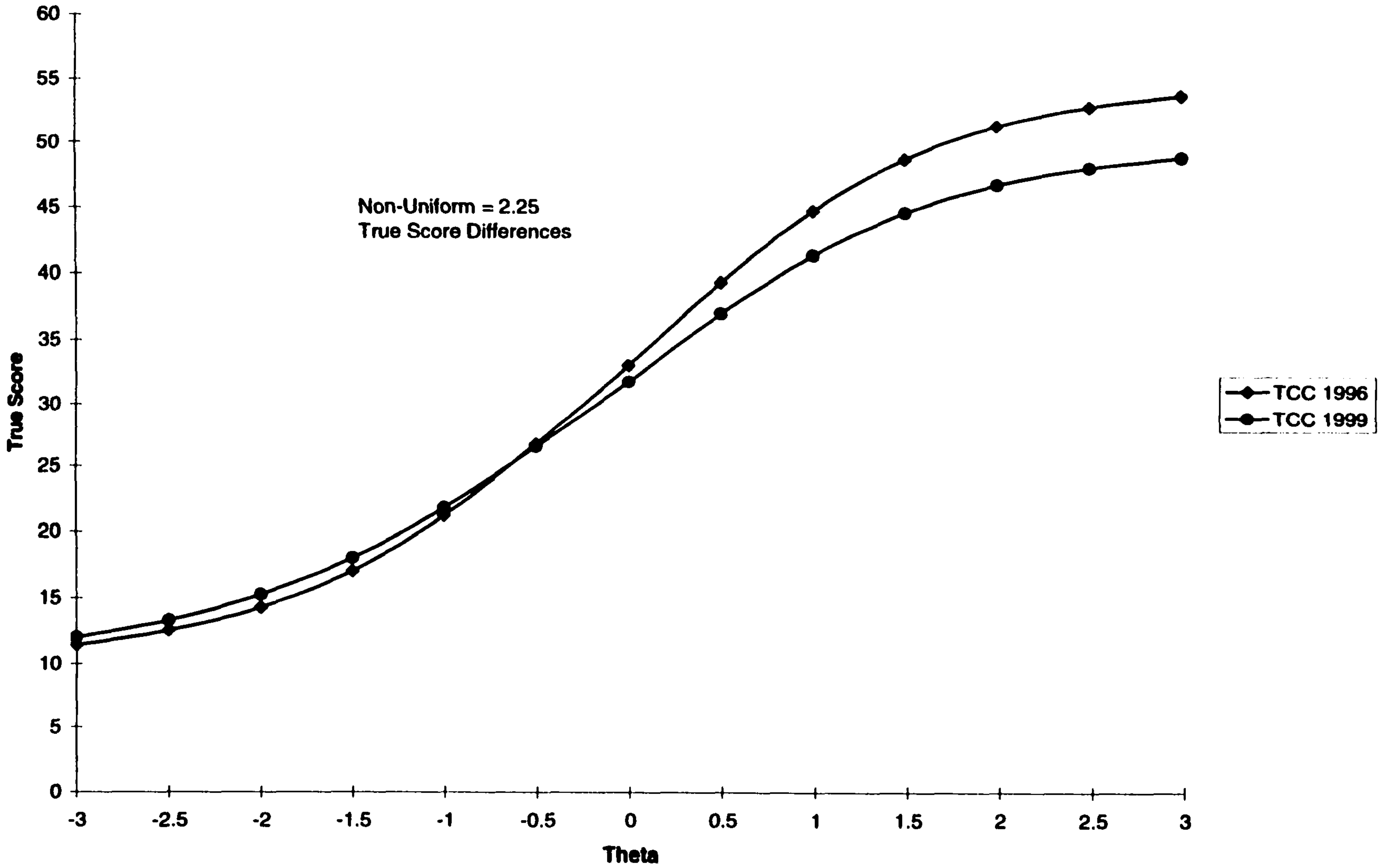


Figure 14. Two Test Characteristic Curves for the Average Squared Difference.

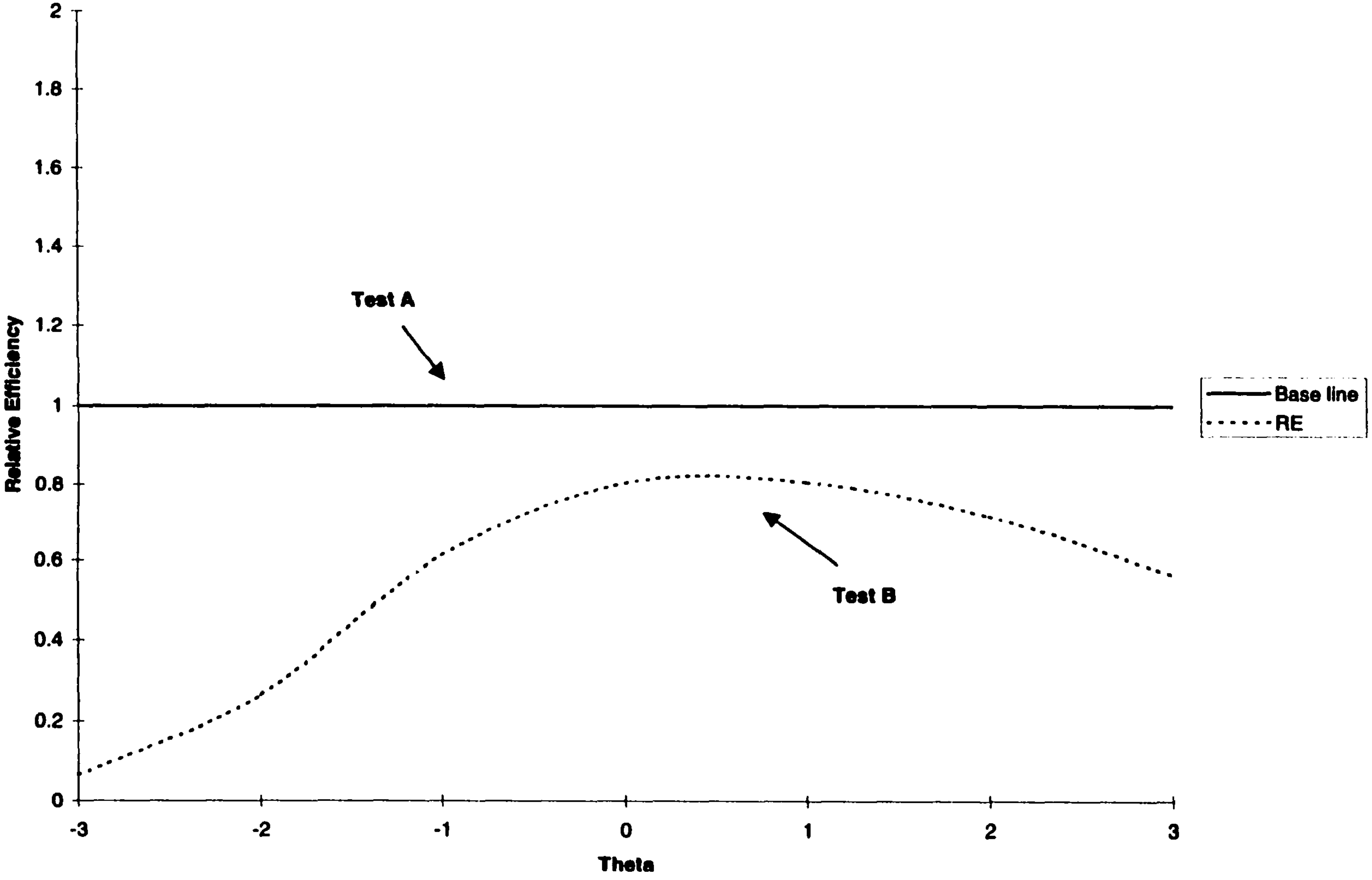


Figure 15. The Relative Efficiency between Test A and Test B.

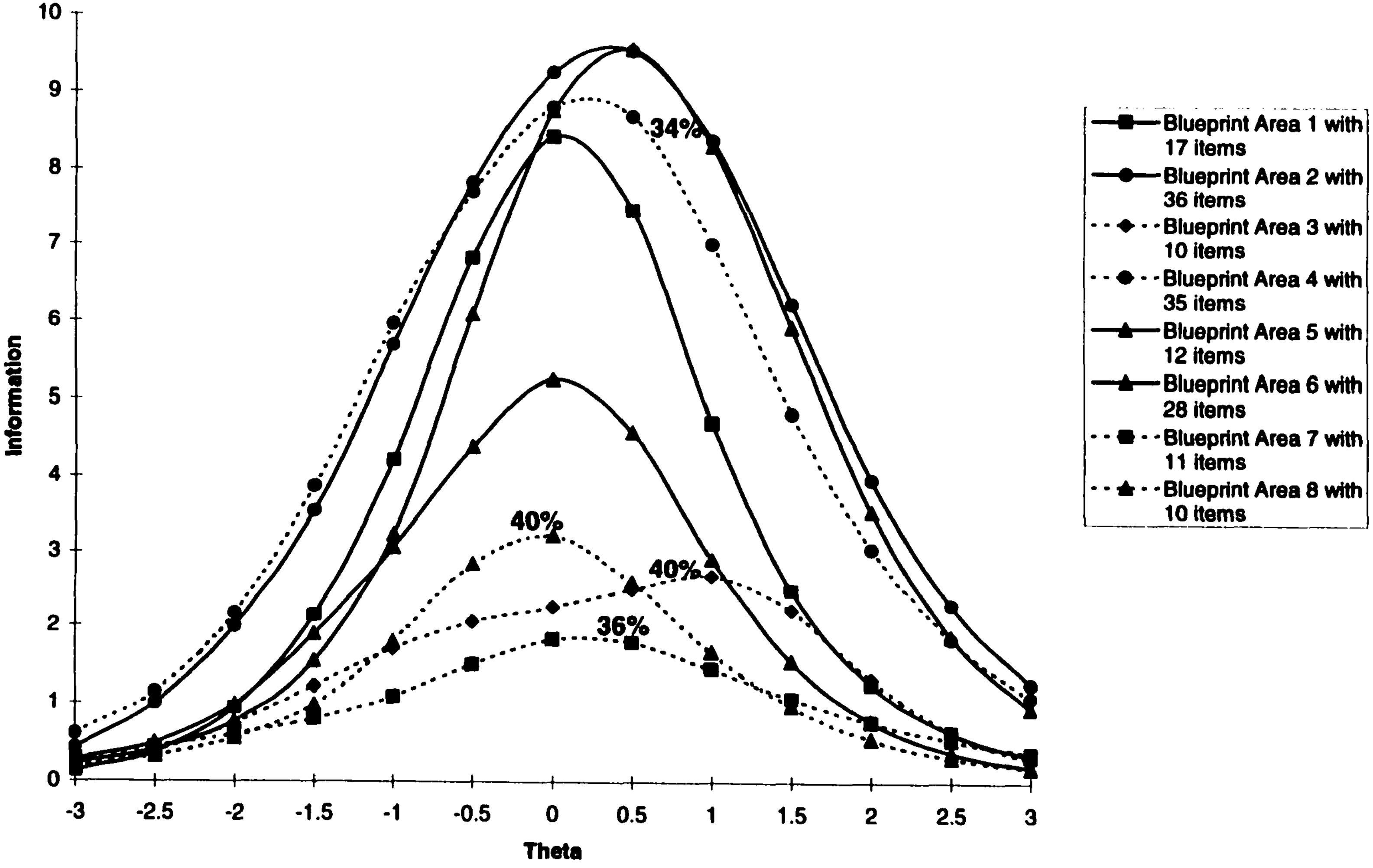


Figure 16. The Achievement Item Bank Information by Blueprint Area.

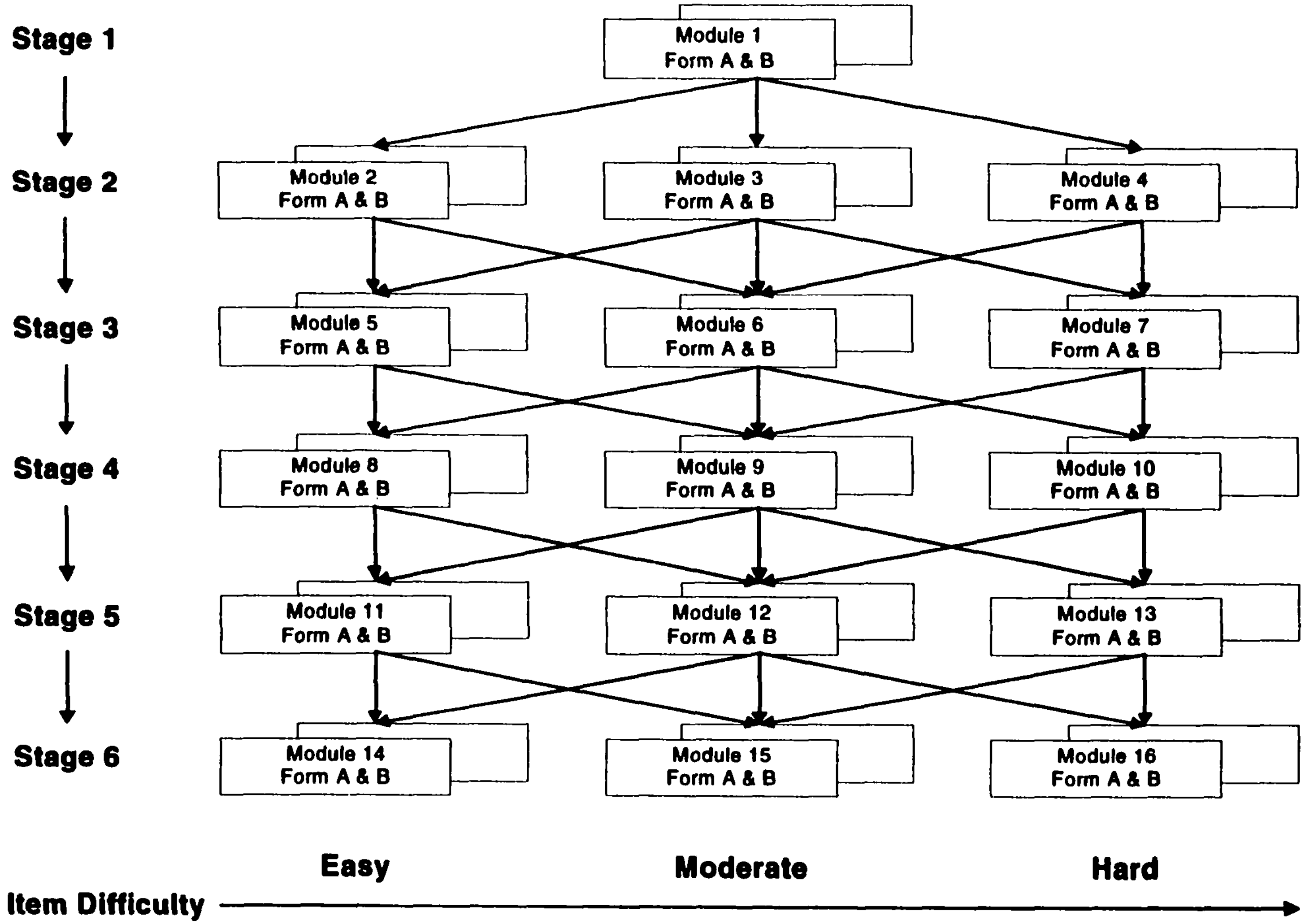


Figure 17. A Six Stage Parallel CAST Model with 32 Modules.

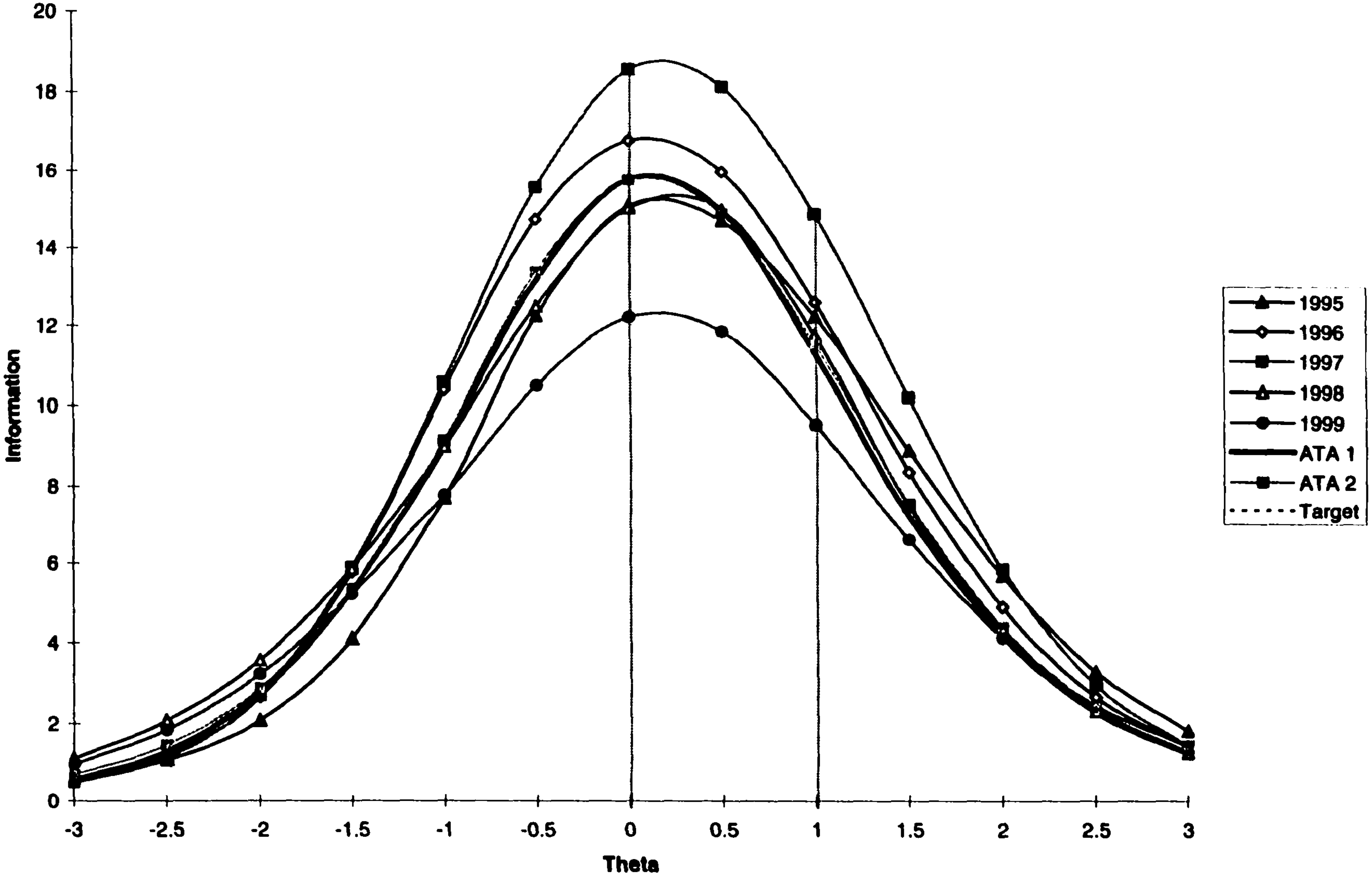


Figure 18. The Information Functions for both the ATA forms and the Original Handcrafted Forms across Years.

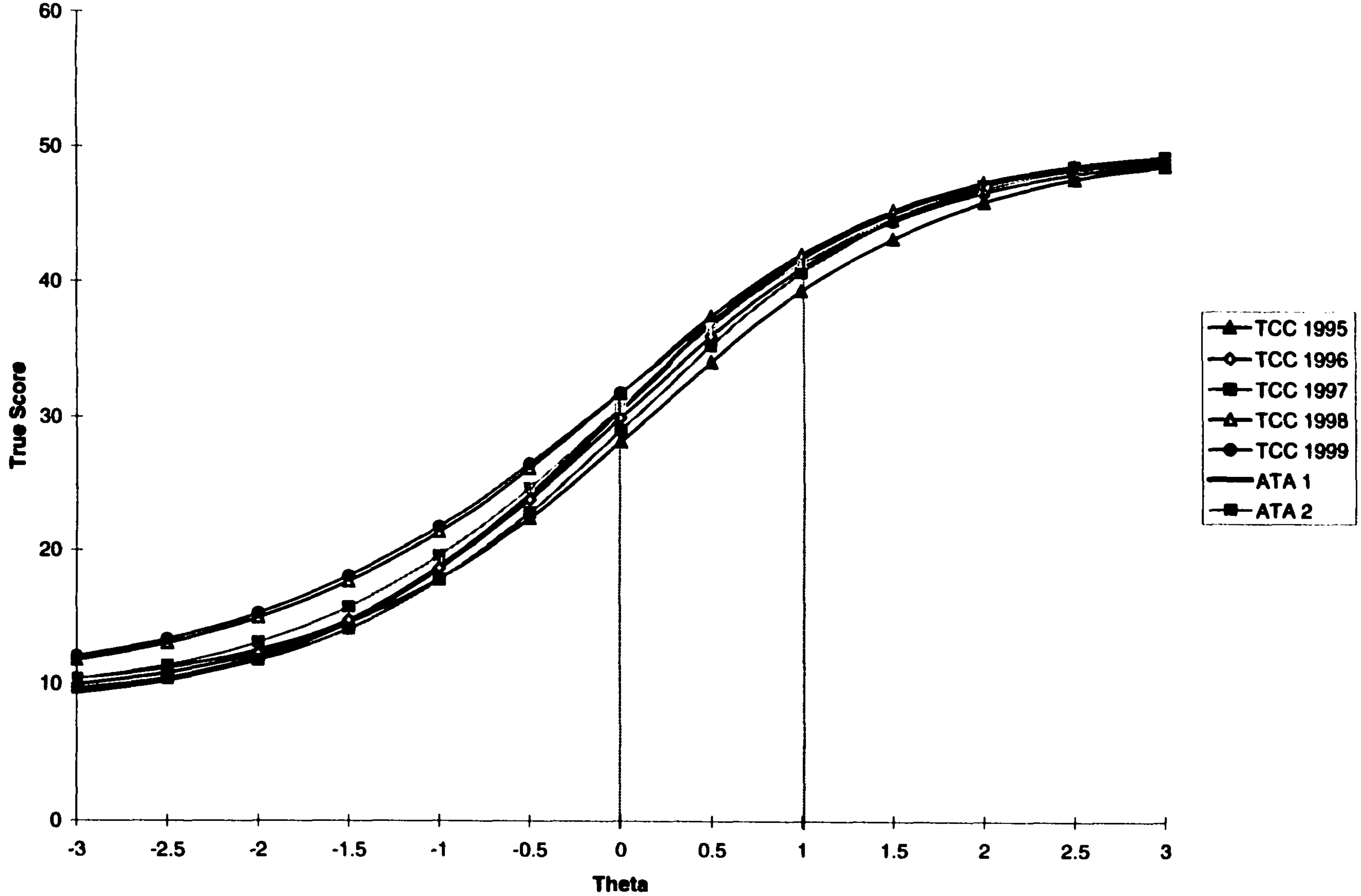
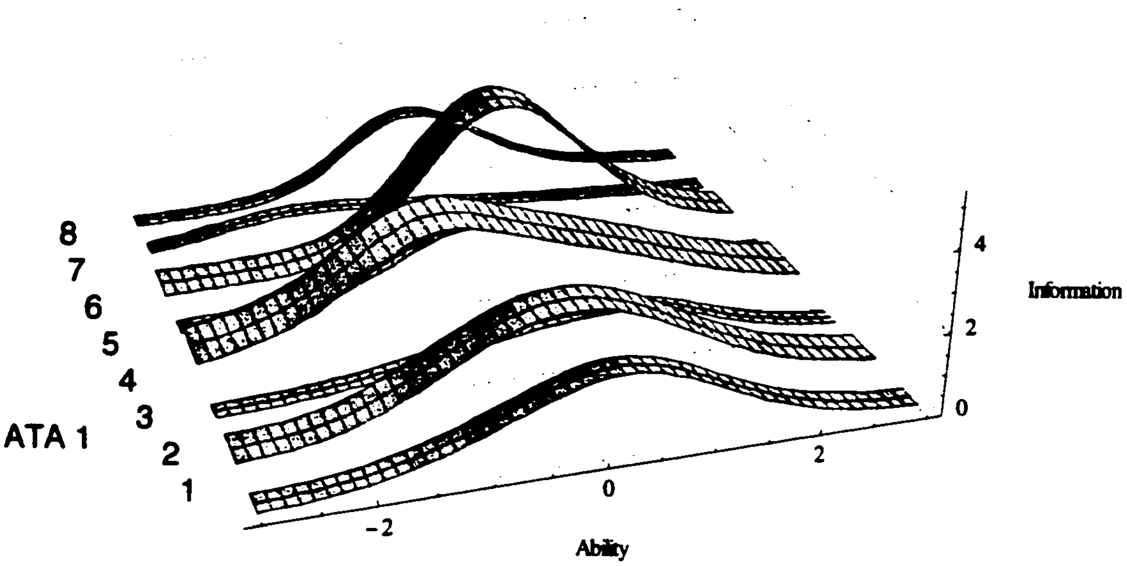


Figure 19. The TCCs for both the ATA forms and the original handcrafted forms from 1995 to 1999.



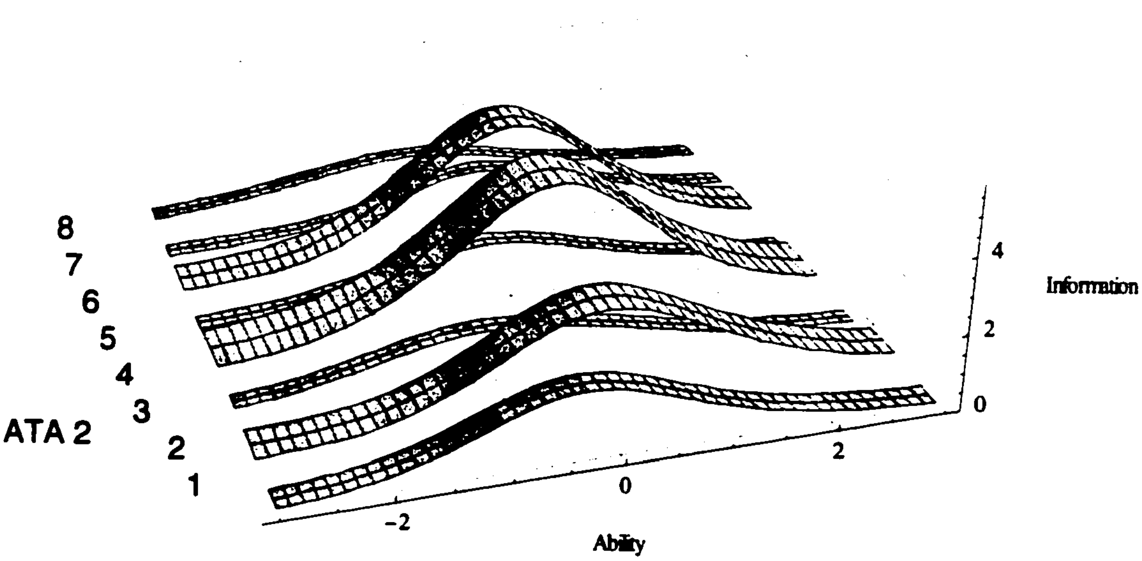


Figure 20. These are the ATA 1 and 2 Forms from the Achievement Item Bank by Blueprint .area.

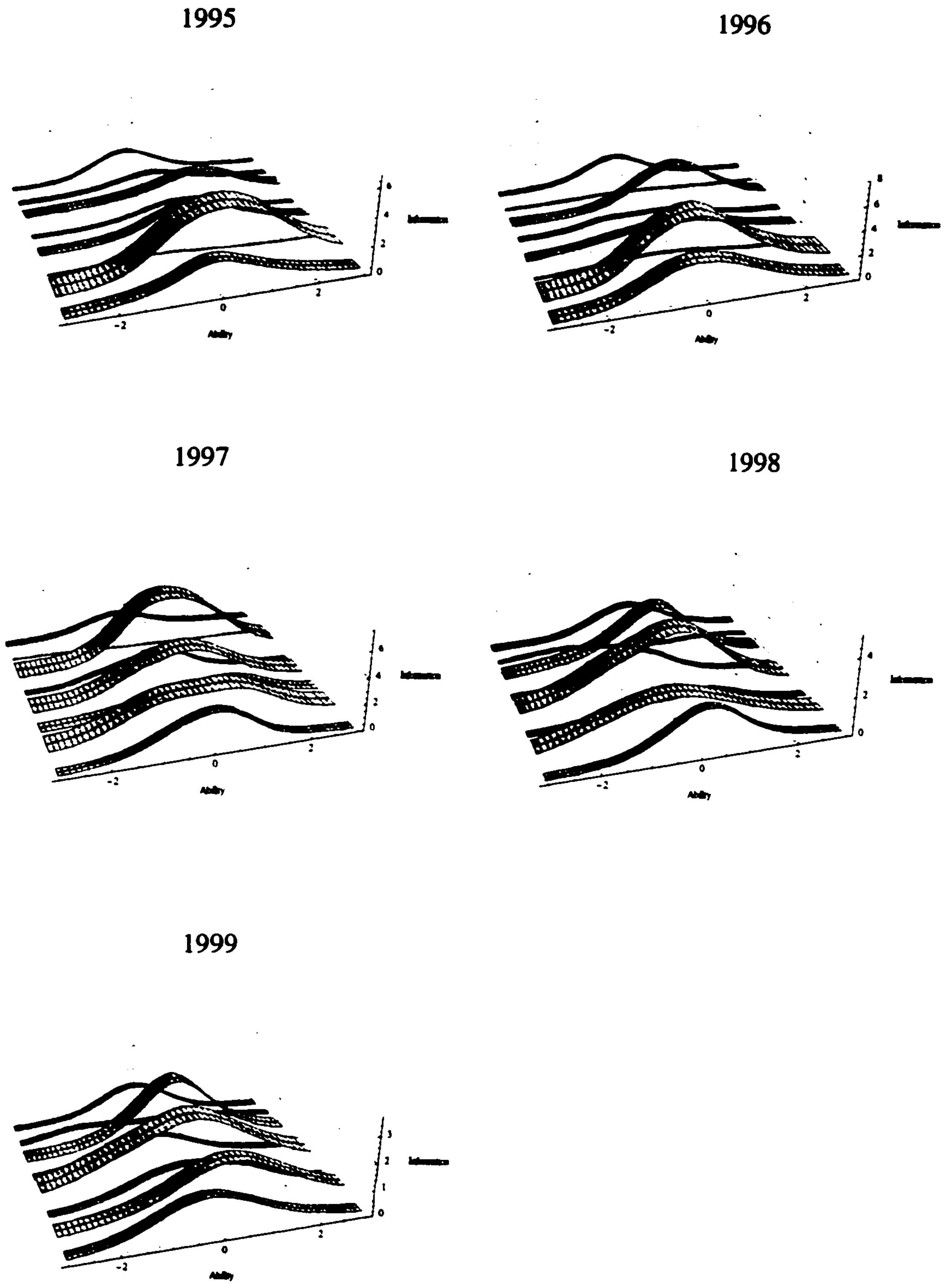


Figure 21. Blueprint Maps for the 1995 to 1999 Grade 9 Achievement Tests.

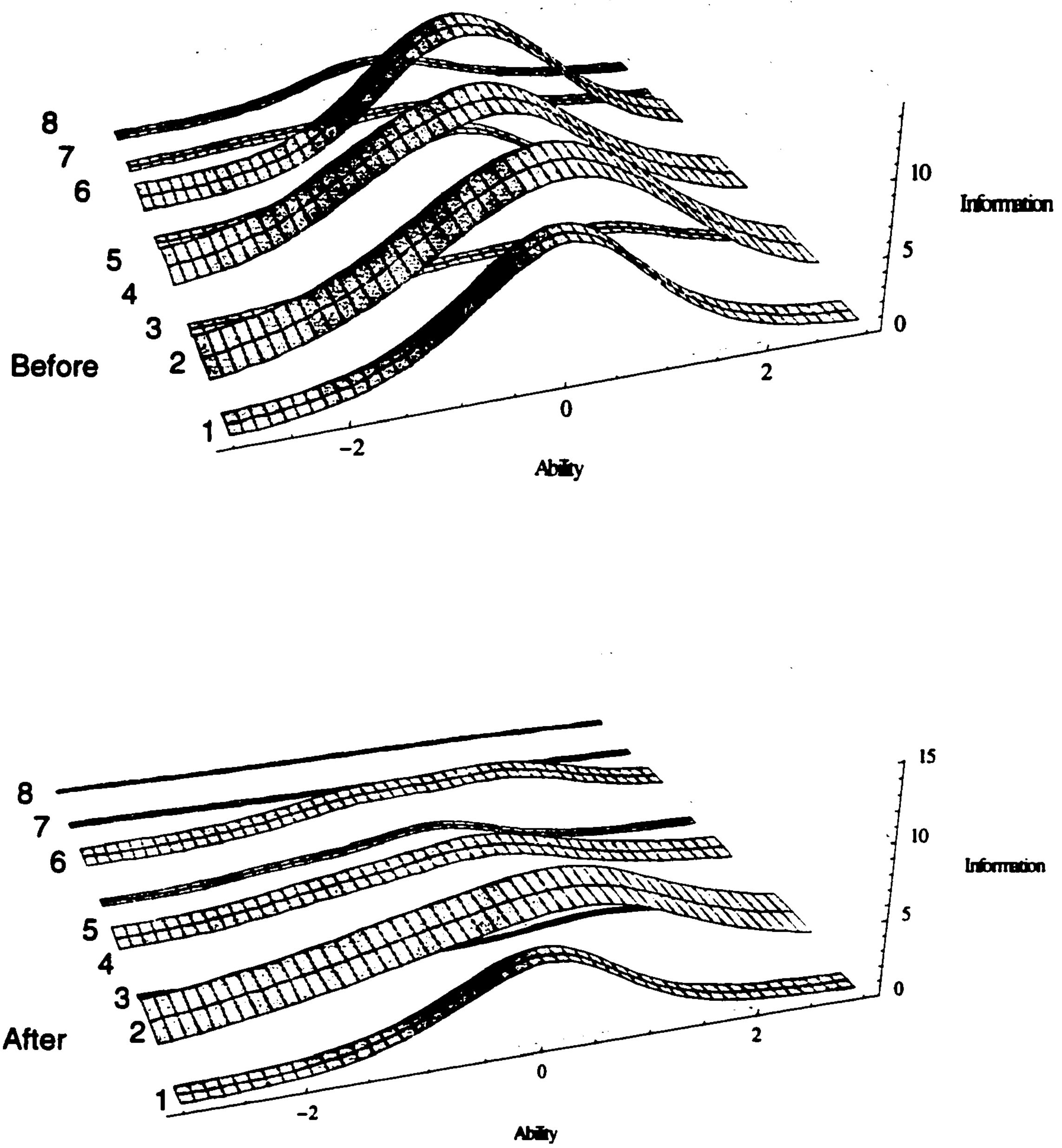


Figure 22. An Achievement Item Bank Consisting of 159 Items.

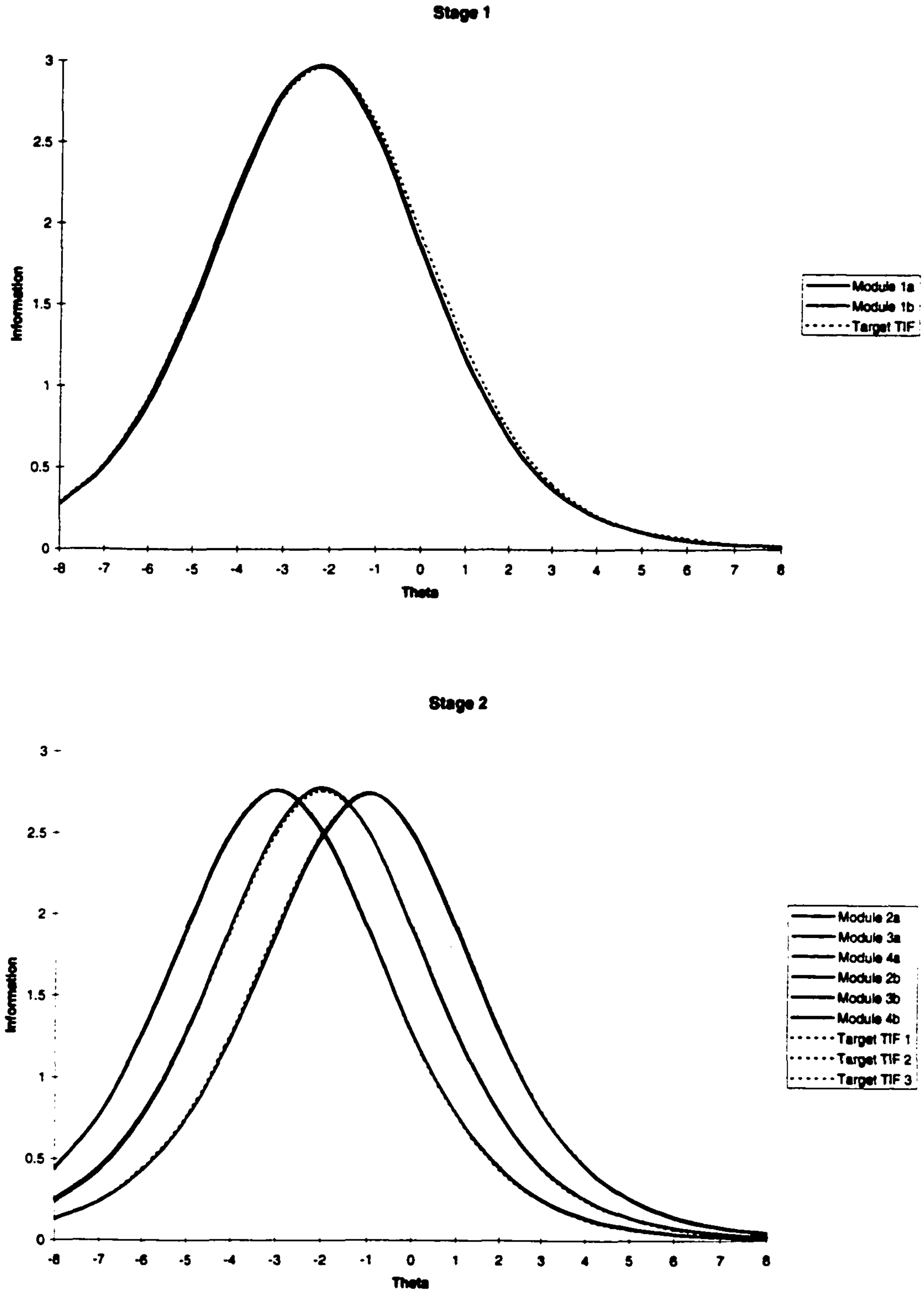
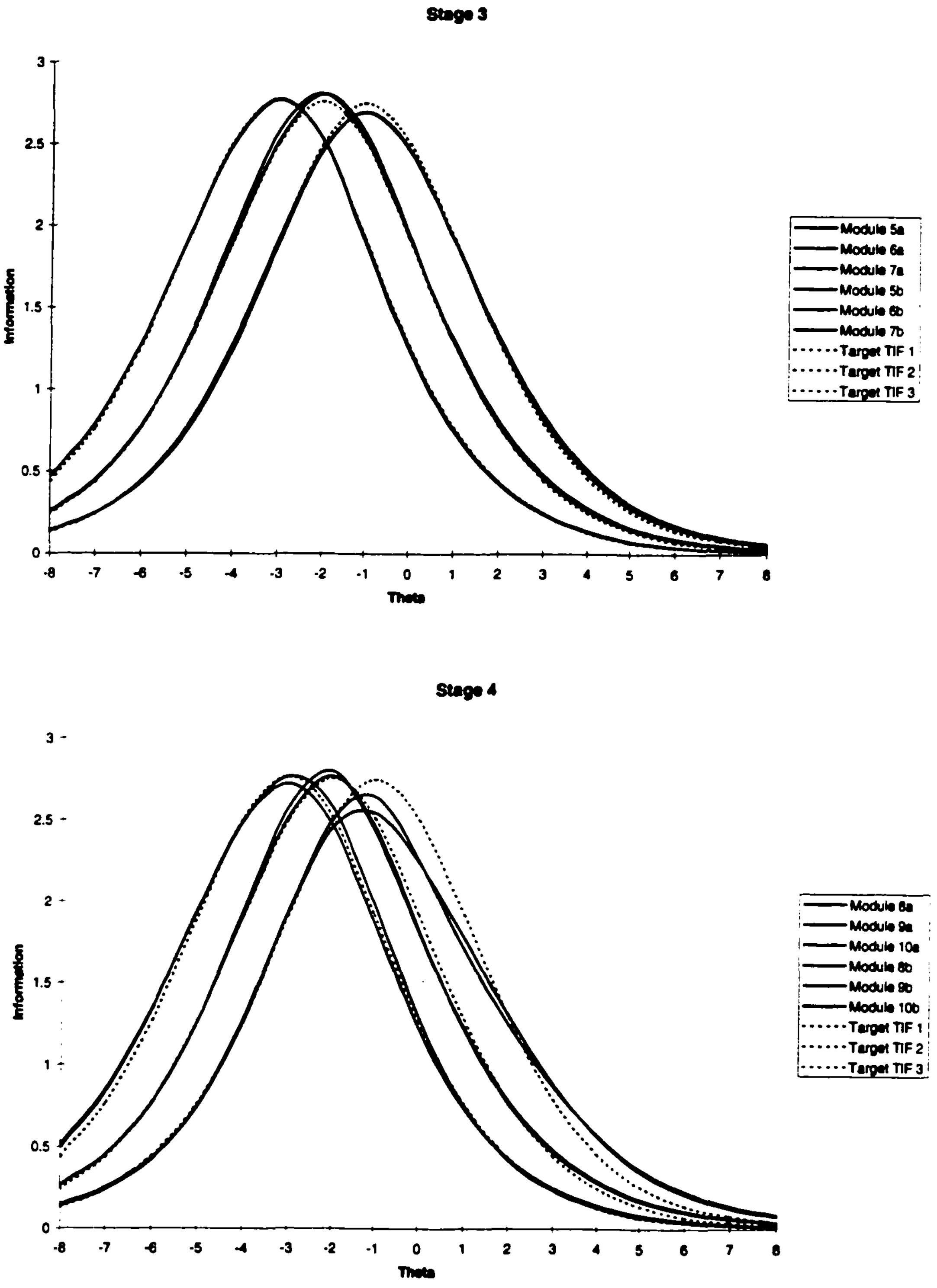


Figure 23a. Stages 1 and 2 for First MCC Panel Assembly.



Stage 3

Stage 4

Figure 23b. Stage 3 and 4 for First MCC Panel Assembly.

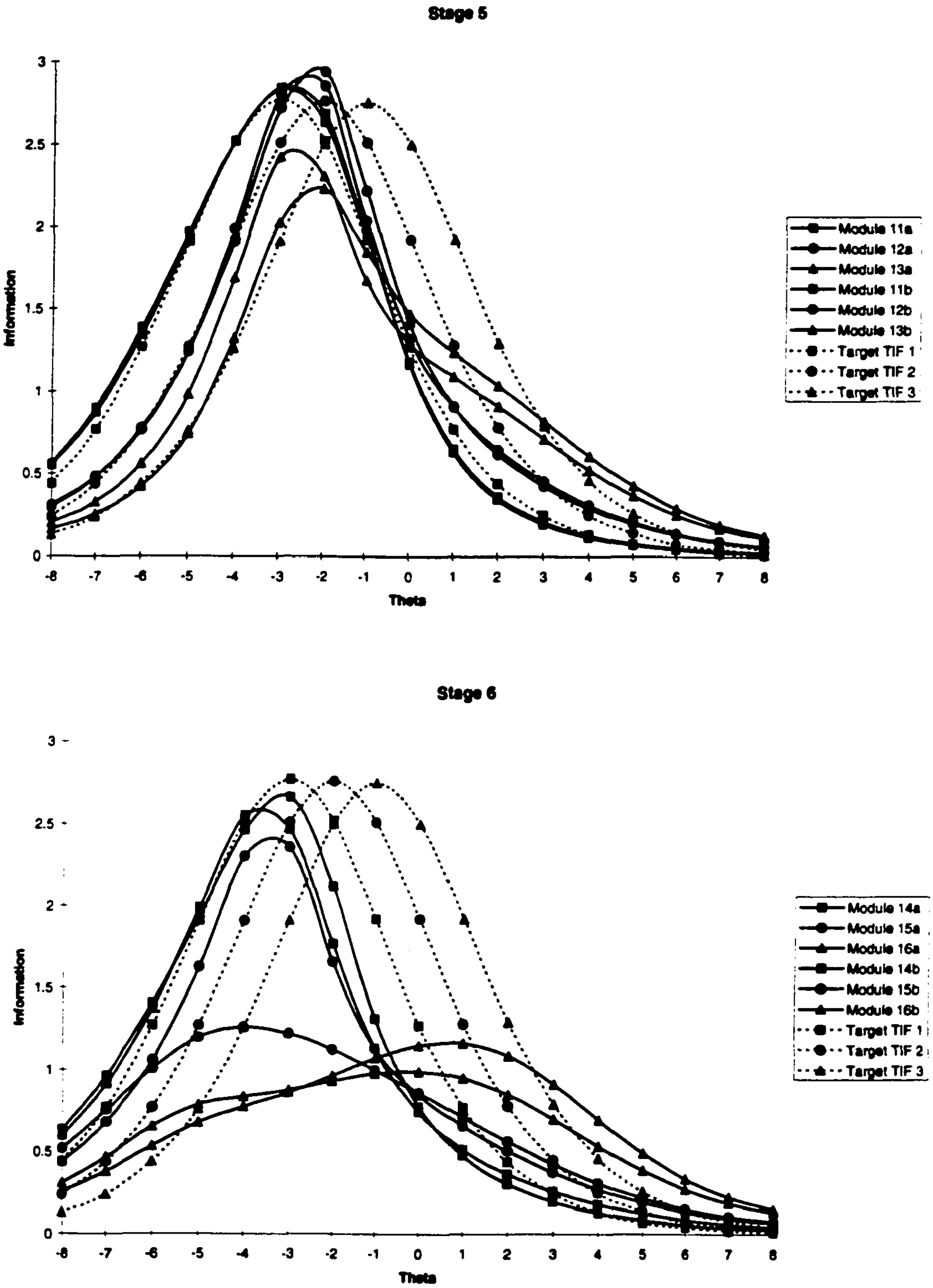
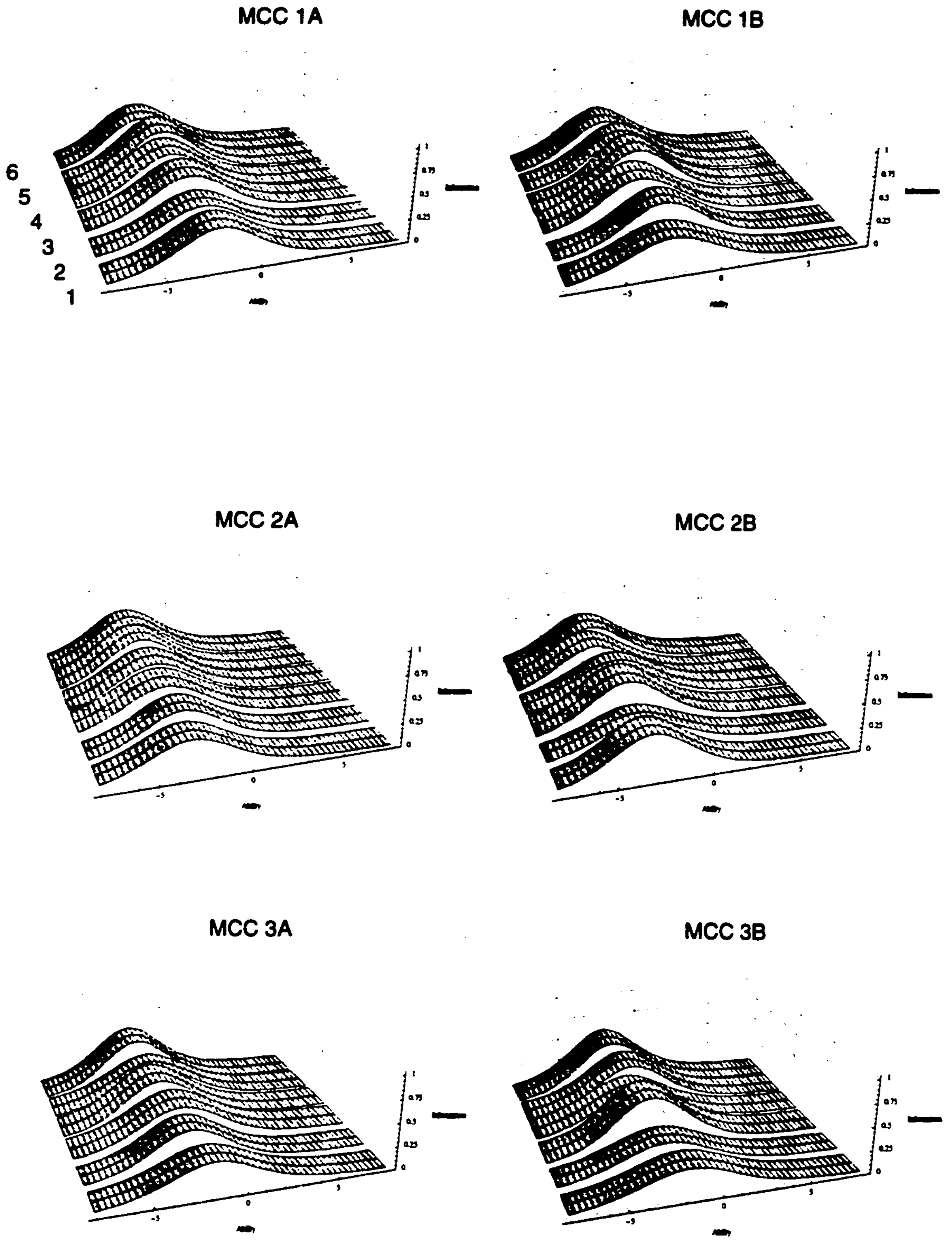


Figure 23c. Stage 5 and 6 for First MCC Panel Assembly.



MCC 1A

MCC 1B

MCC 2A

MCC 2B

MCC 3A

MCC 3B

Figure 24a. Modules 1, 2, and 3 for First MCC Panel Assembly.

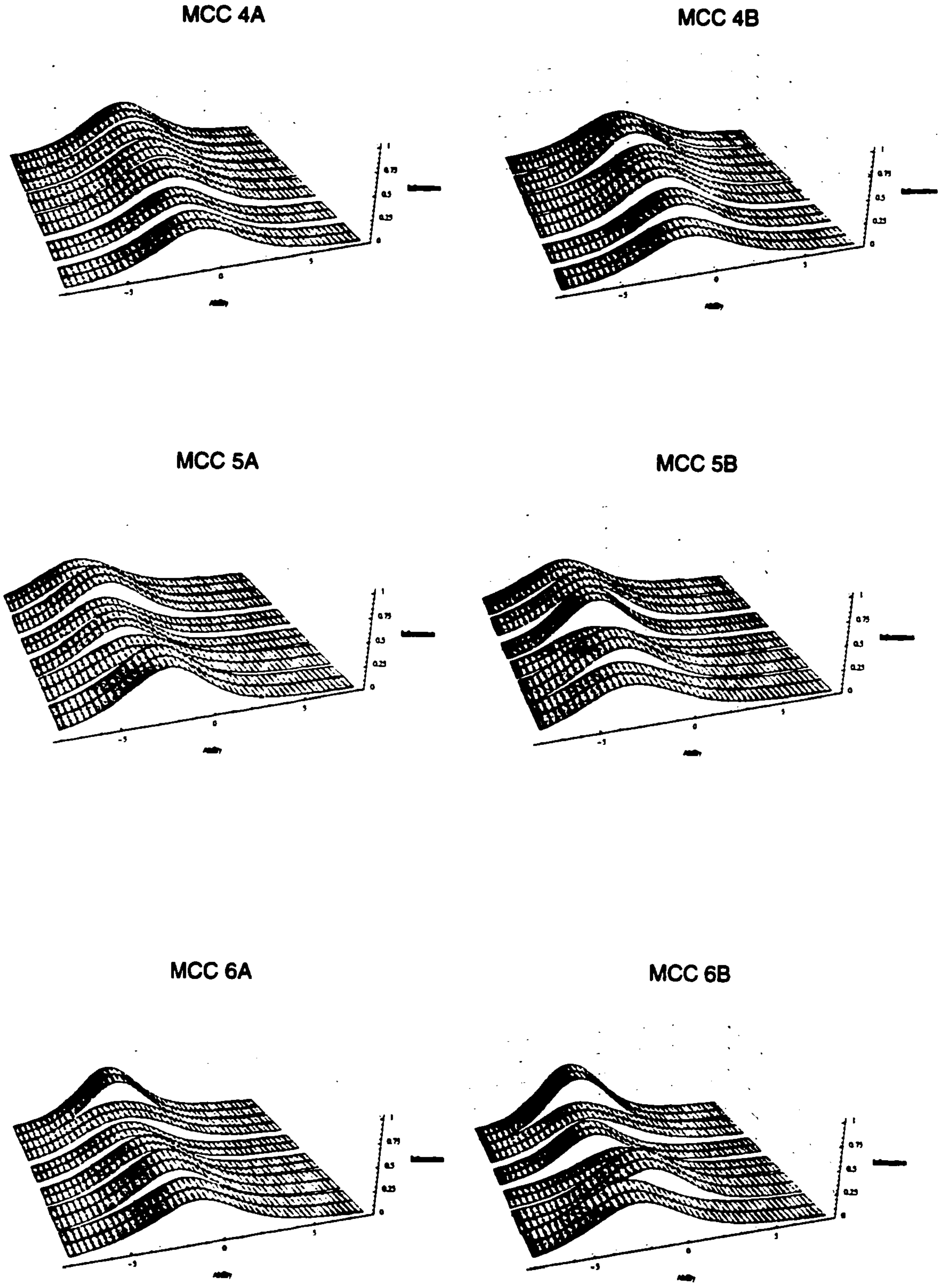
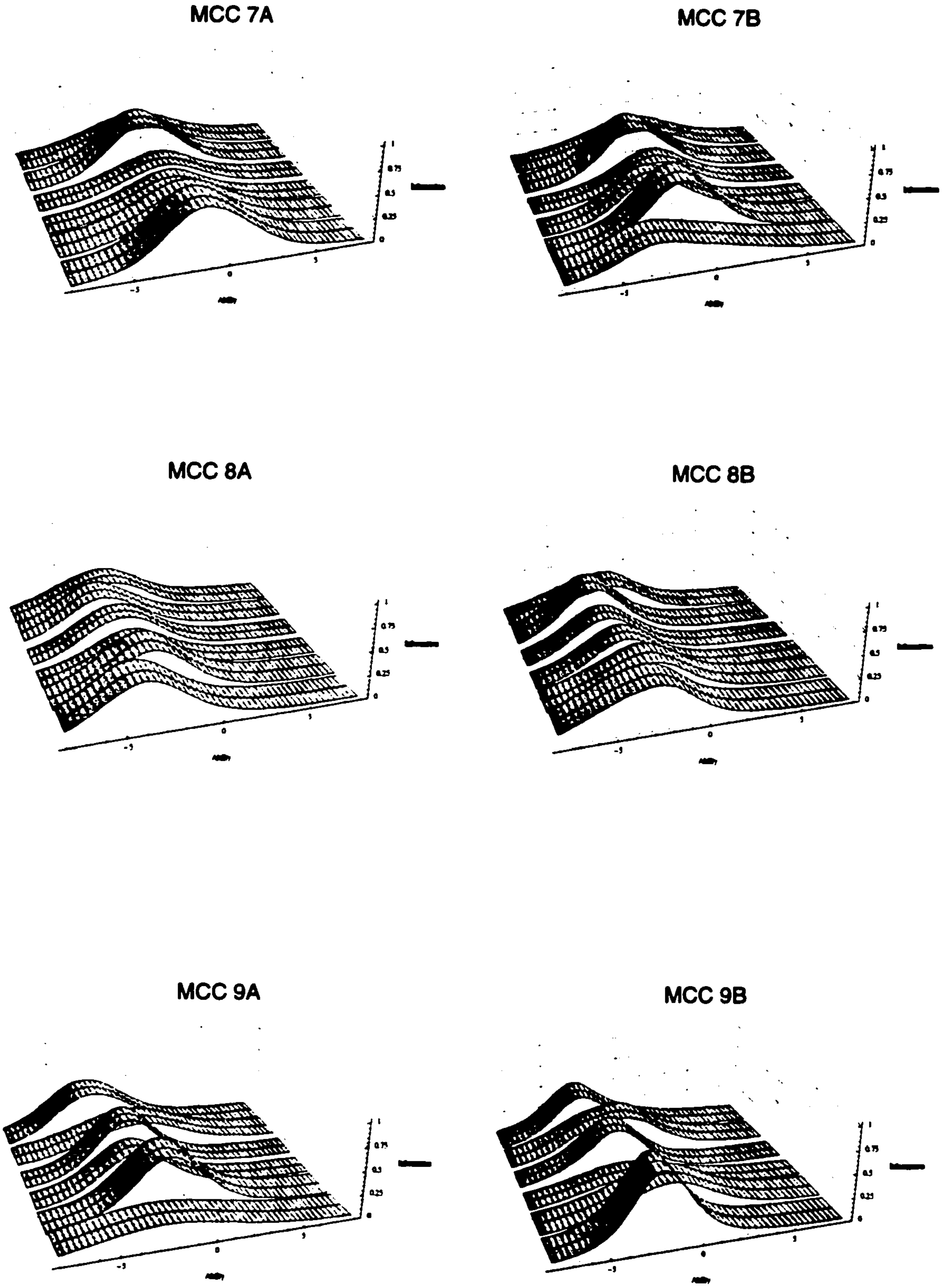


Figure 24b. Modules 4, 5, and 6 for First MCC Panel Assembly.



MCC 7A

MCC 7B

MCC 8A

MCC 8B

MCC 9A

MCC 9B

Figure 24c. Modules 7, 8, and 9 for First MCC Panel Assembly.

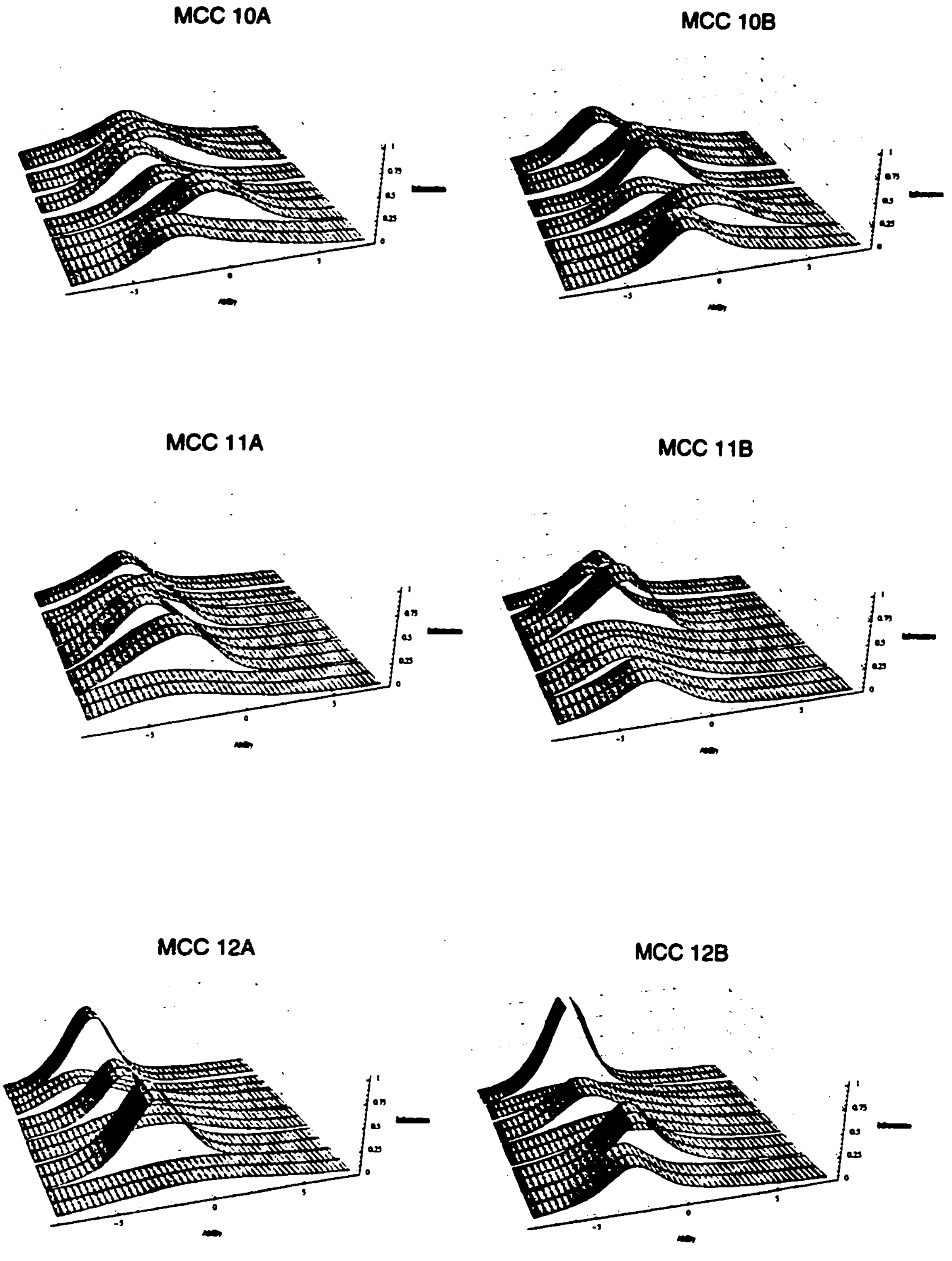


Figure 24d. Modules 10, 11, and 12 for First MCC Panel Assembly.

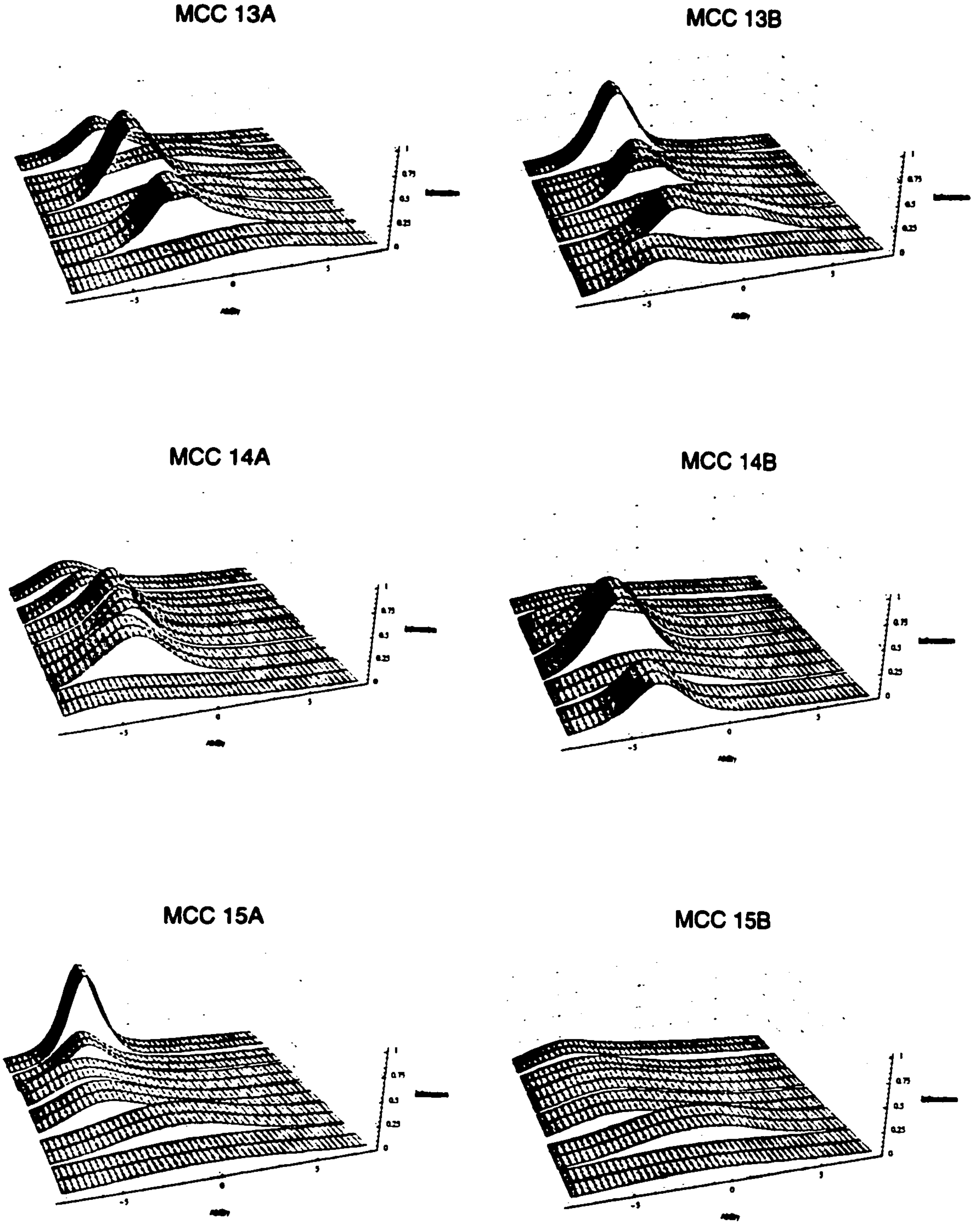
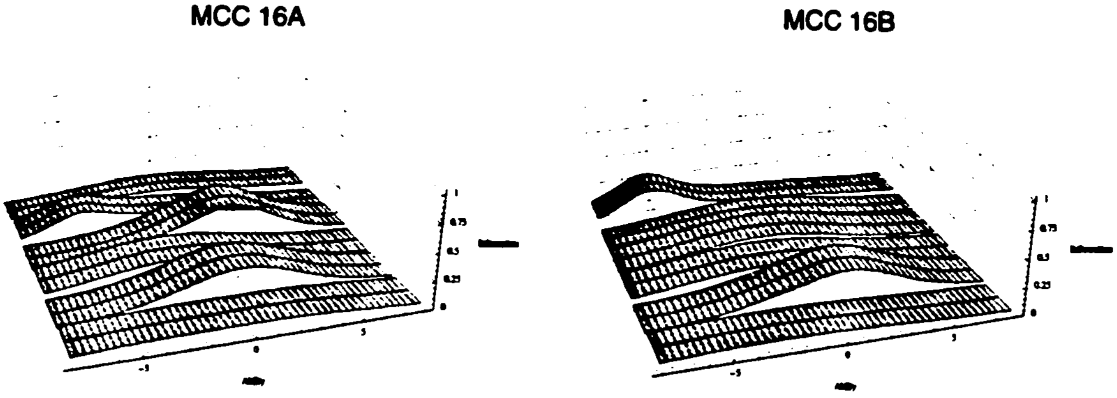


Figure 24e. Modules 13, 14, and 15 for First MCC Panel Assembly.


MCC 16A

MCC 16B

Figure 24f. Module 16 for First MCC Panel Assembly.

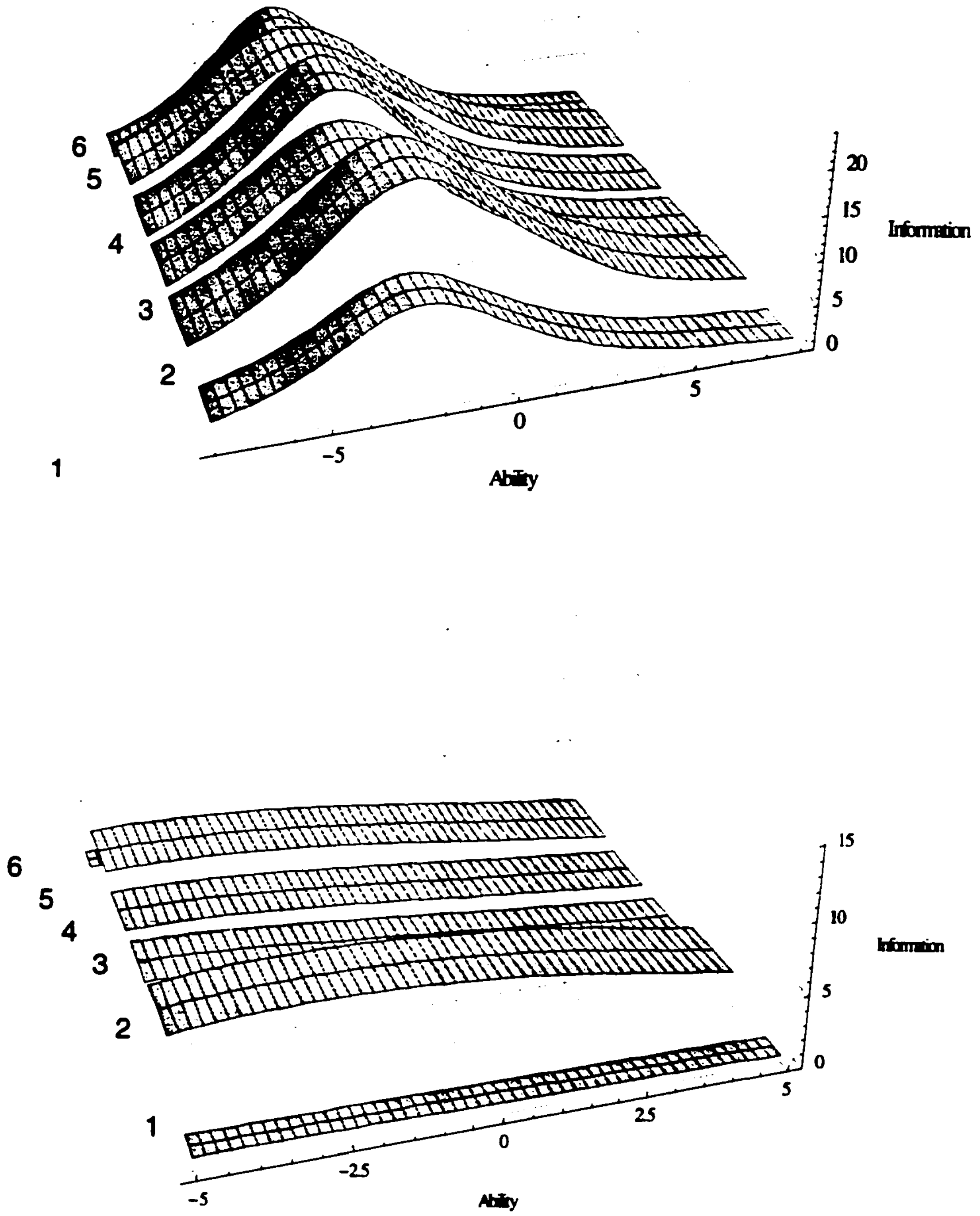


Figure 25. MCC Item Bank Information Blueprints.

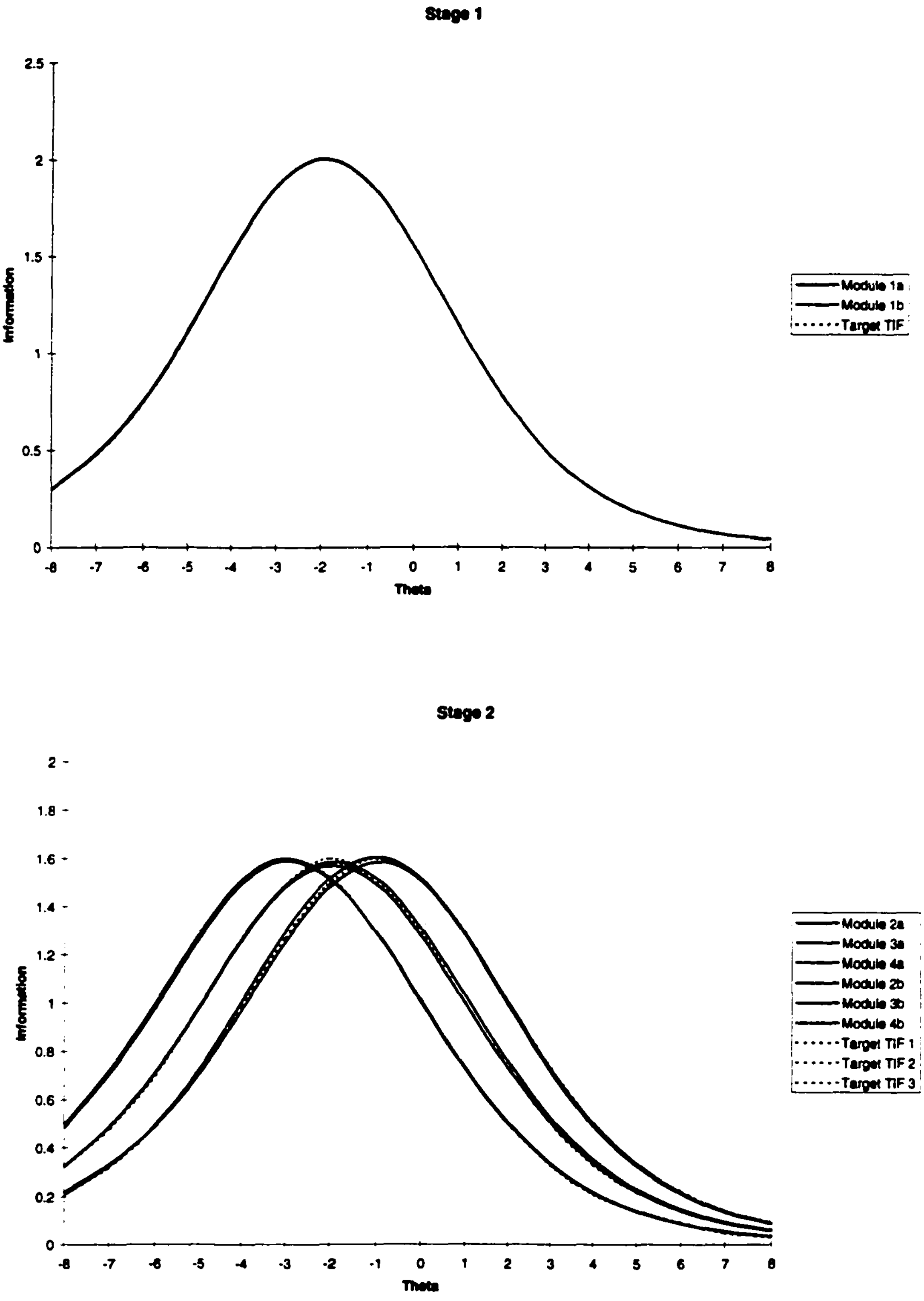


Figure 26a. Stage 1 and 2 for Second MCC Panel Assembly.

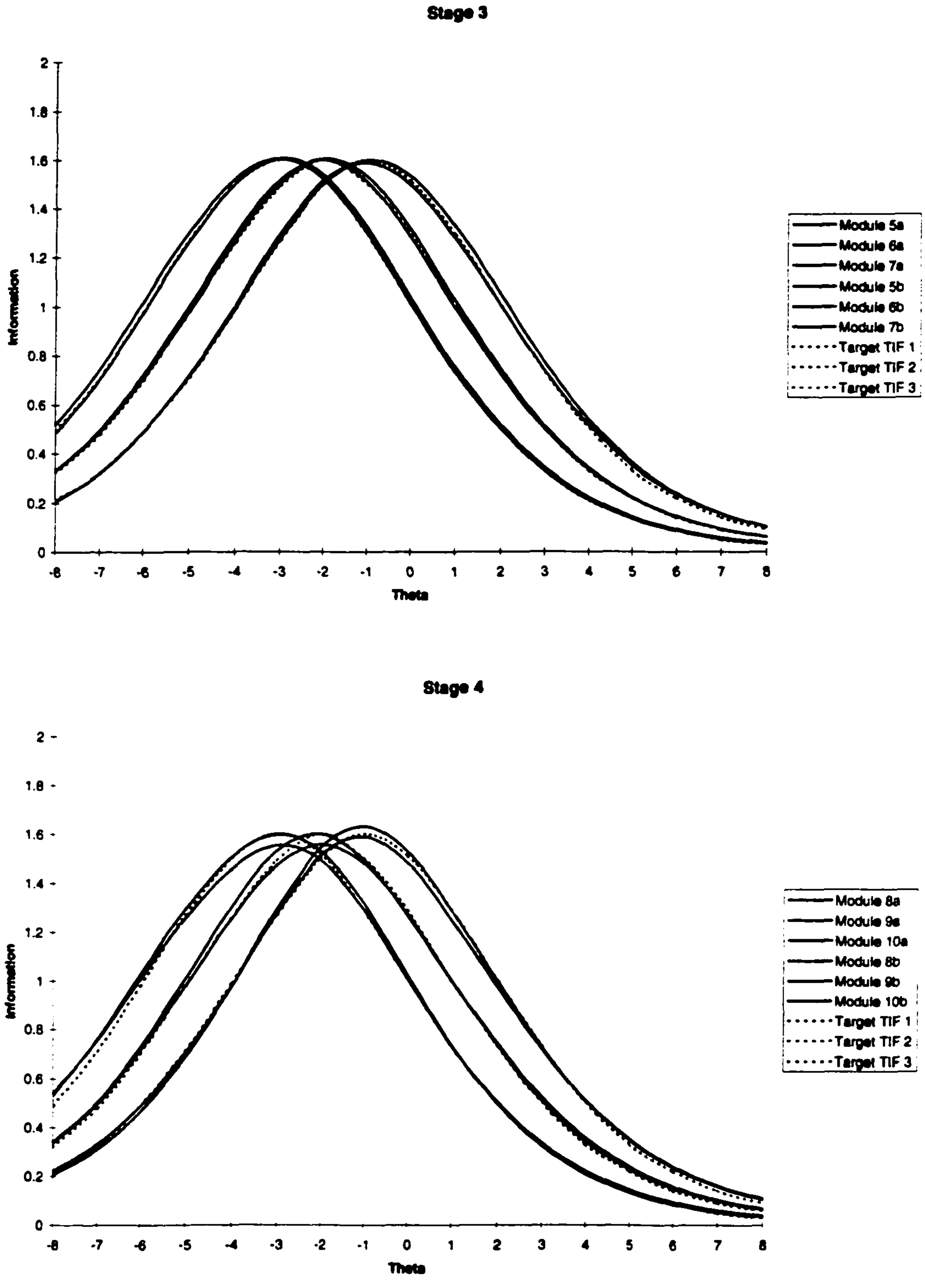


Figure 26b. Stage 3 and 4 for Second MCC Panel Assembly.

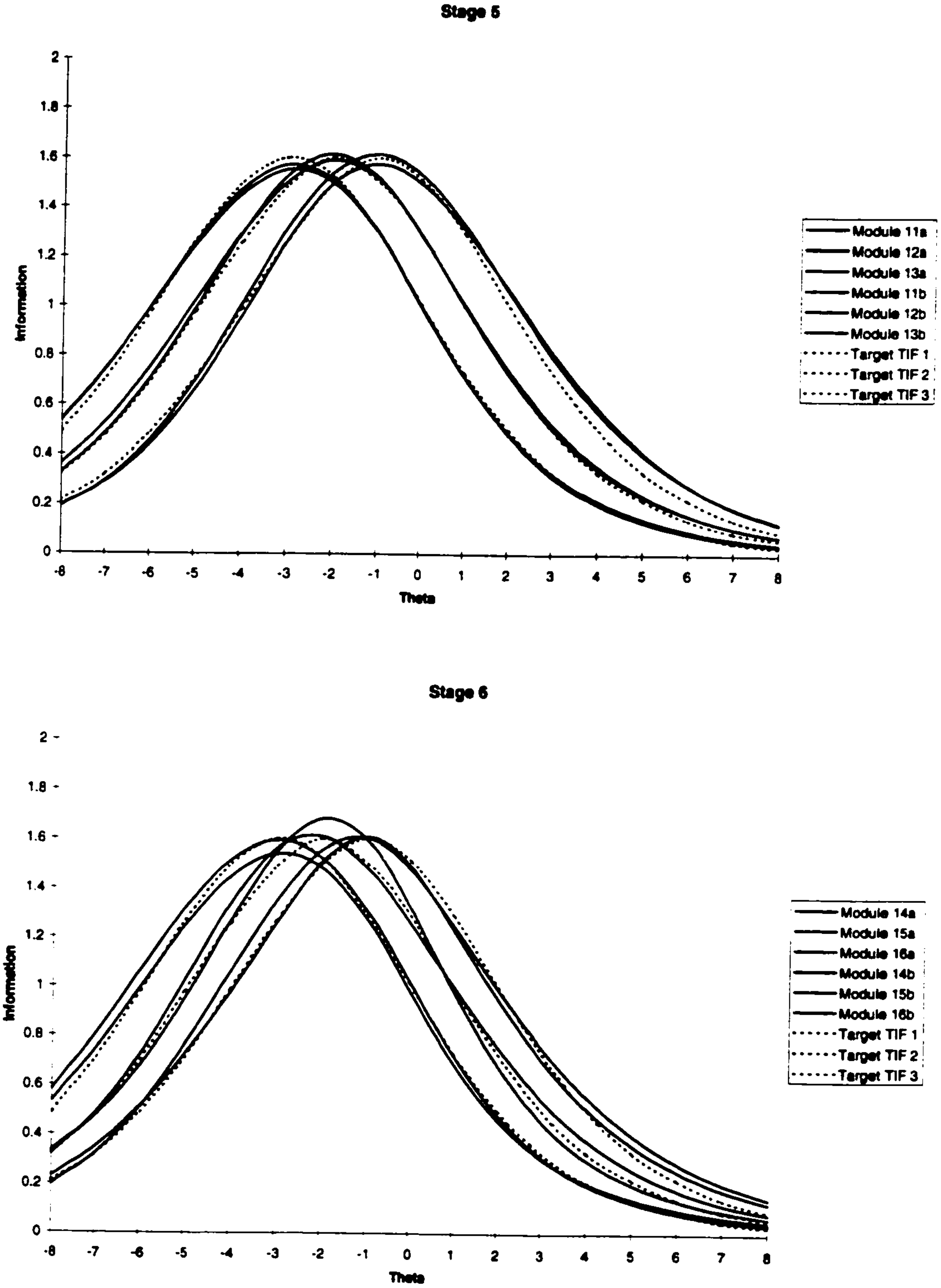


Figure 26c. Stage 5 and 6 for Second MCC Panel Assembly.

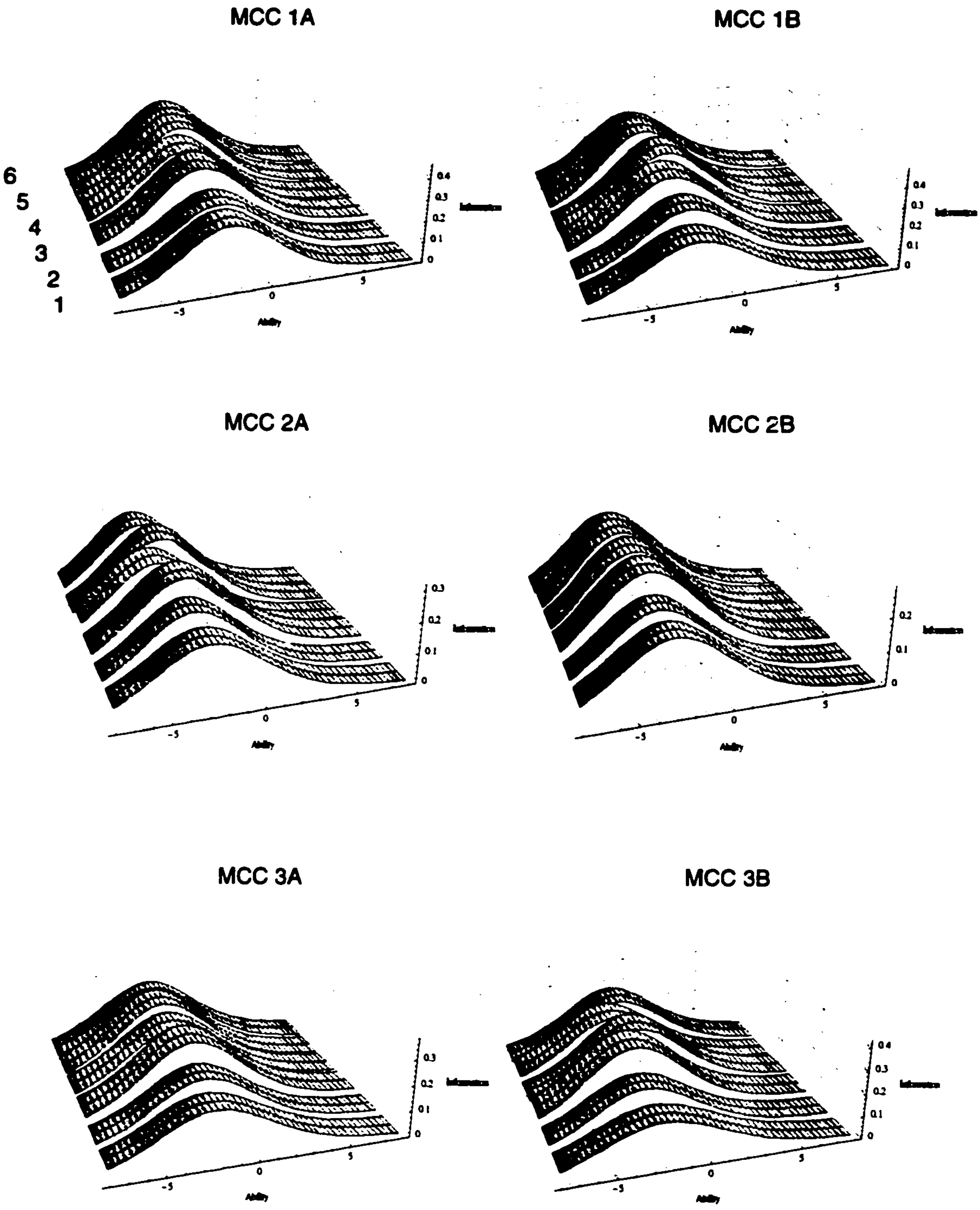


Figure 27a. Modules 1, 2, and 3 for Second MCC Panel Assembly.

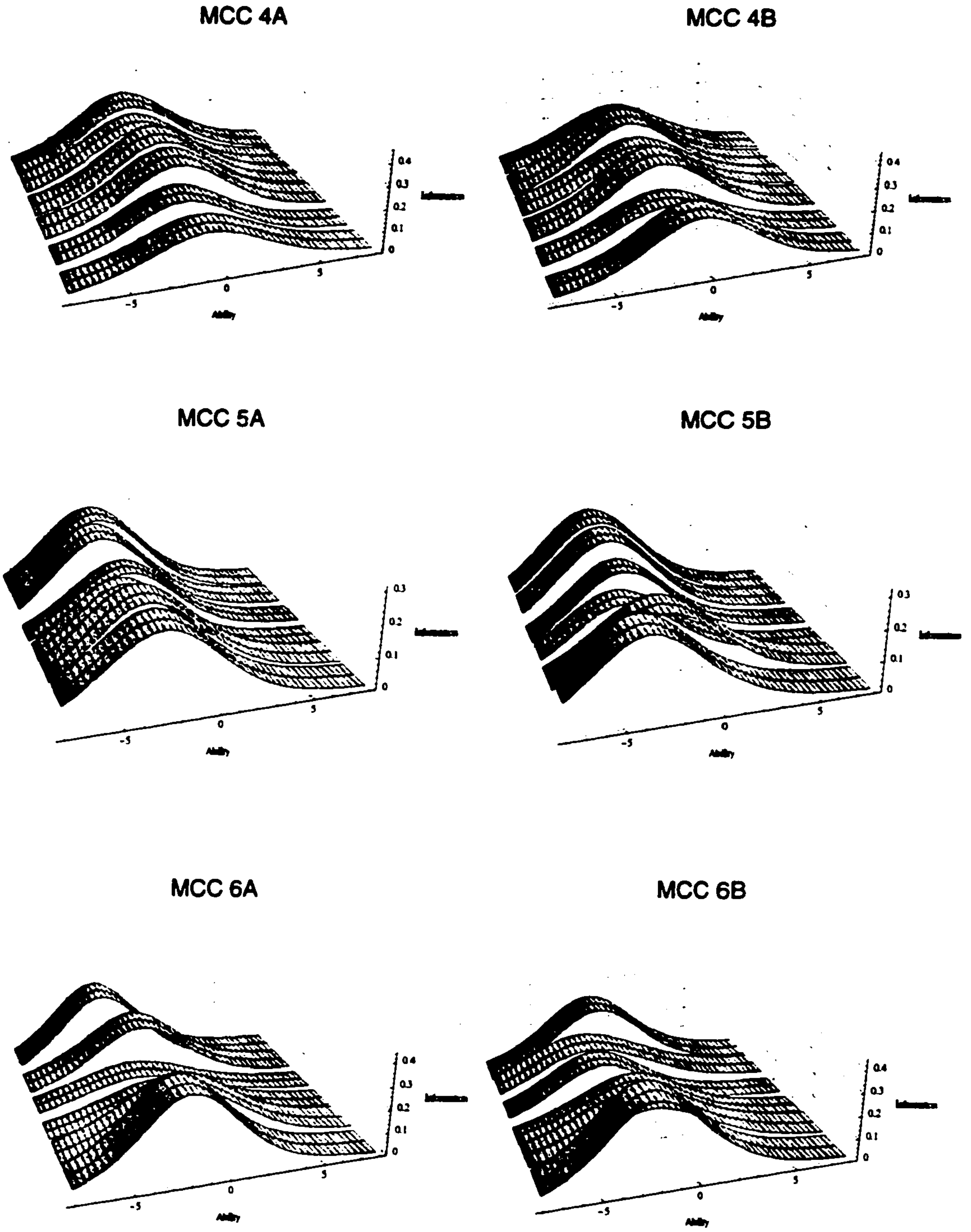


Figure 27b. Modules 4, 5, and 6 for Second MCC Panel Assembly.

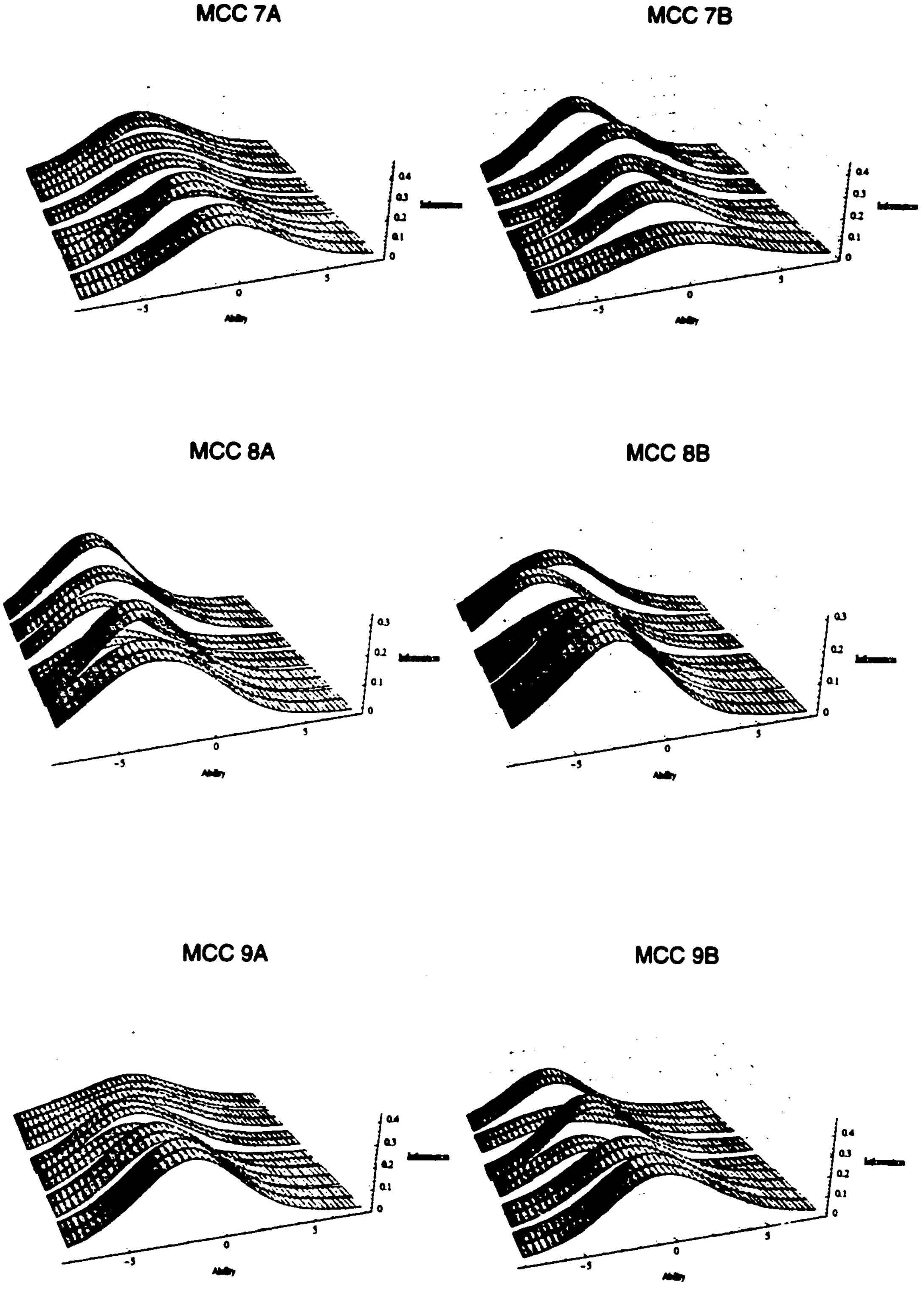
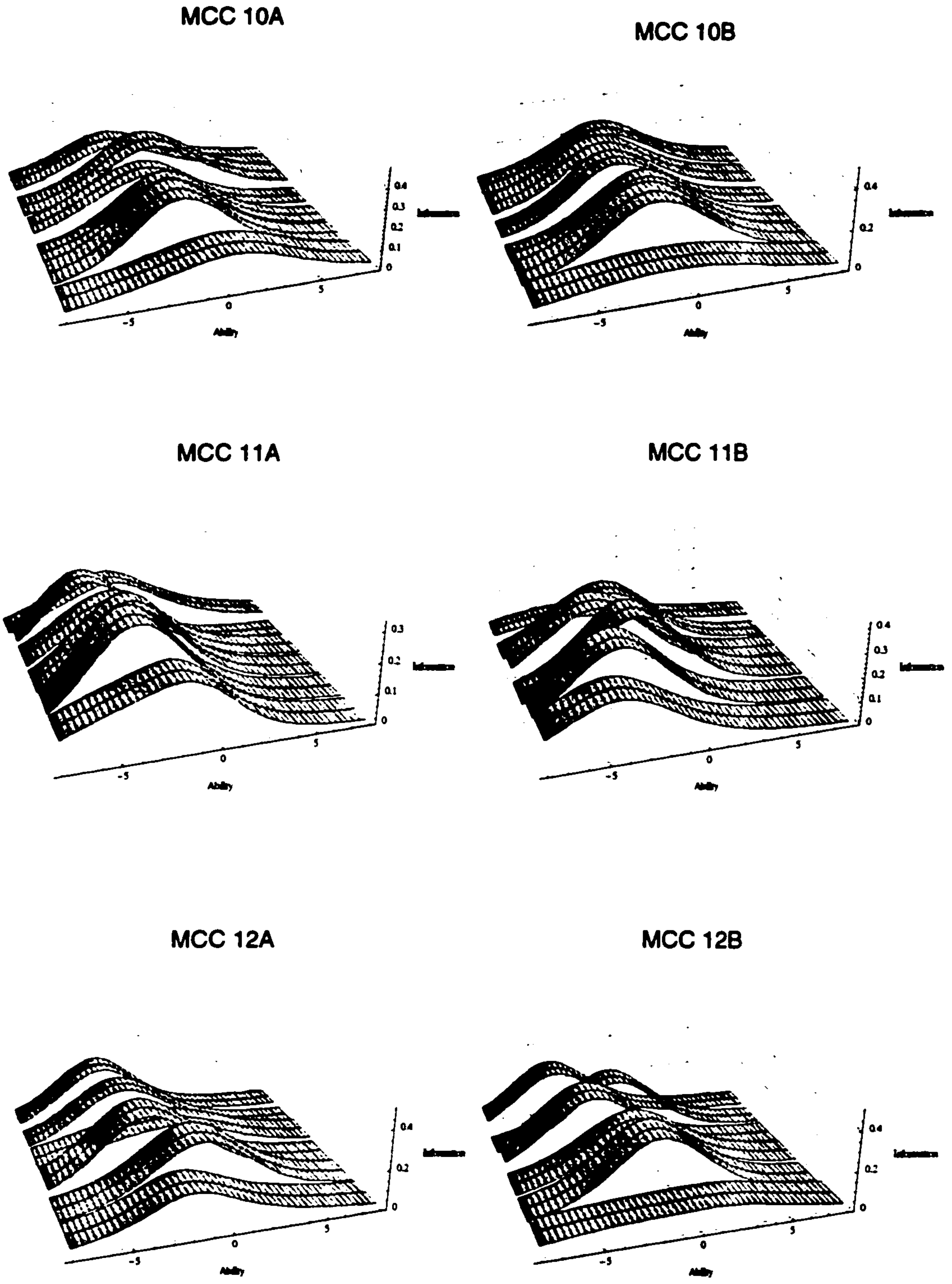


Figure 27c. Modules 7, 8, and 9 for Second MCC Panel Assembly.



MCC 10A

MCC 10B

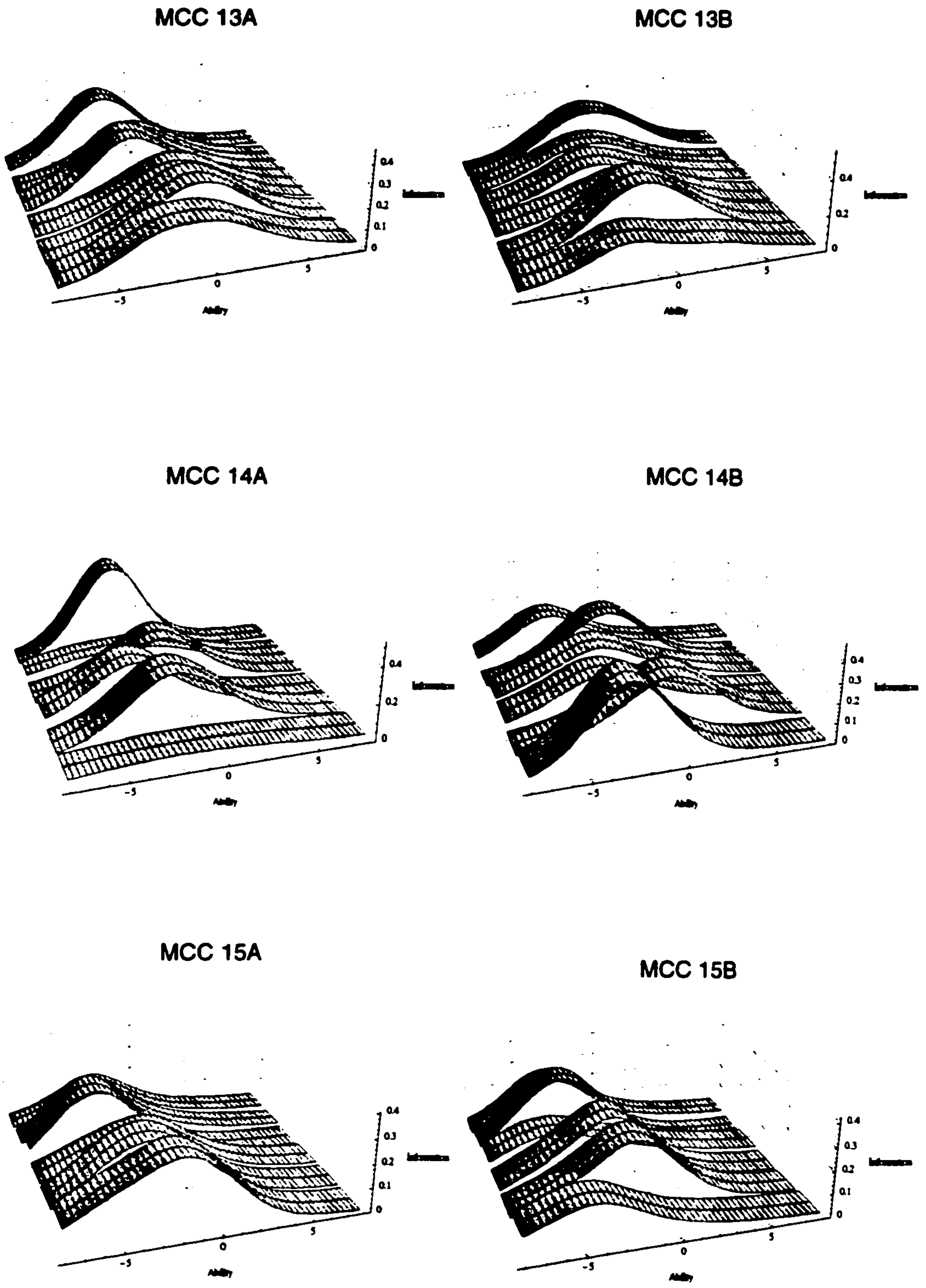
MCC 11A

MCC 11B

MCC 12A

MCC 12B

Figure 27d. Modules 10, 11, and 12 for Second MCC Panel Assembly.



MCC 13A

MCC 13B

MCC 14A

MCC 14B

MCC 15A

MCC 15B

Figure 27e. Modules 13, 14, and 15 for Second MCC Panel Assembly.

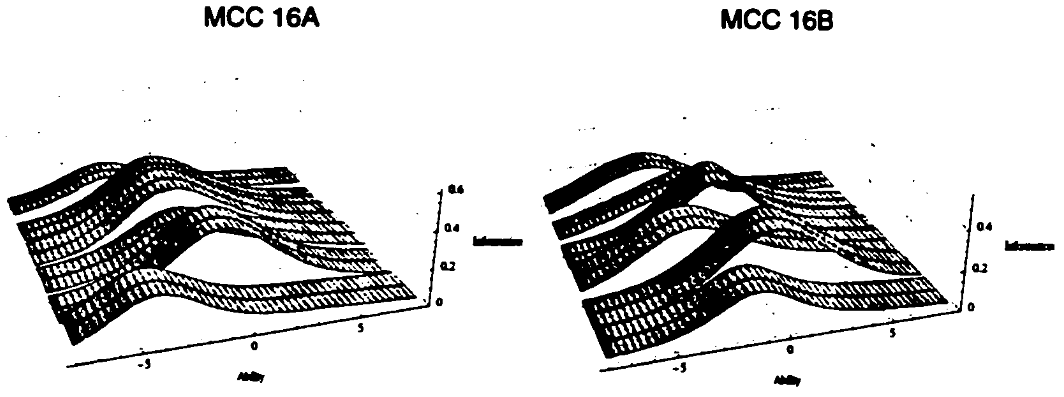


Figure 27f. Module 16 for Second MCC Panel Assembly.

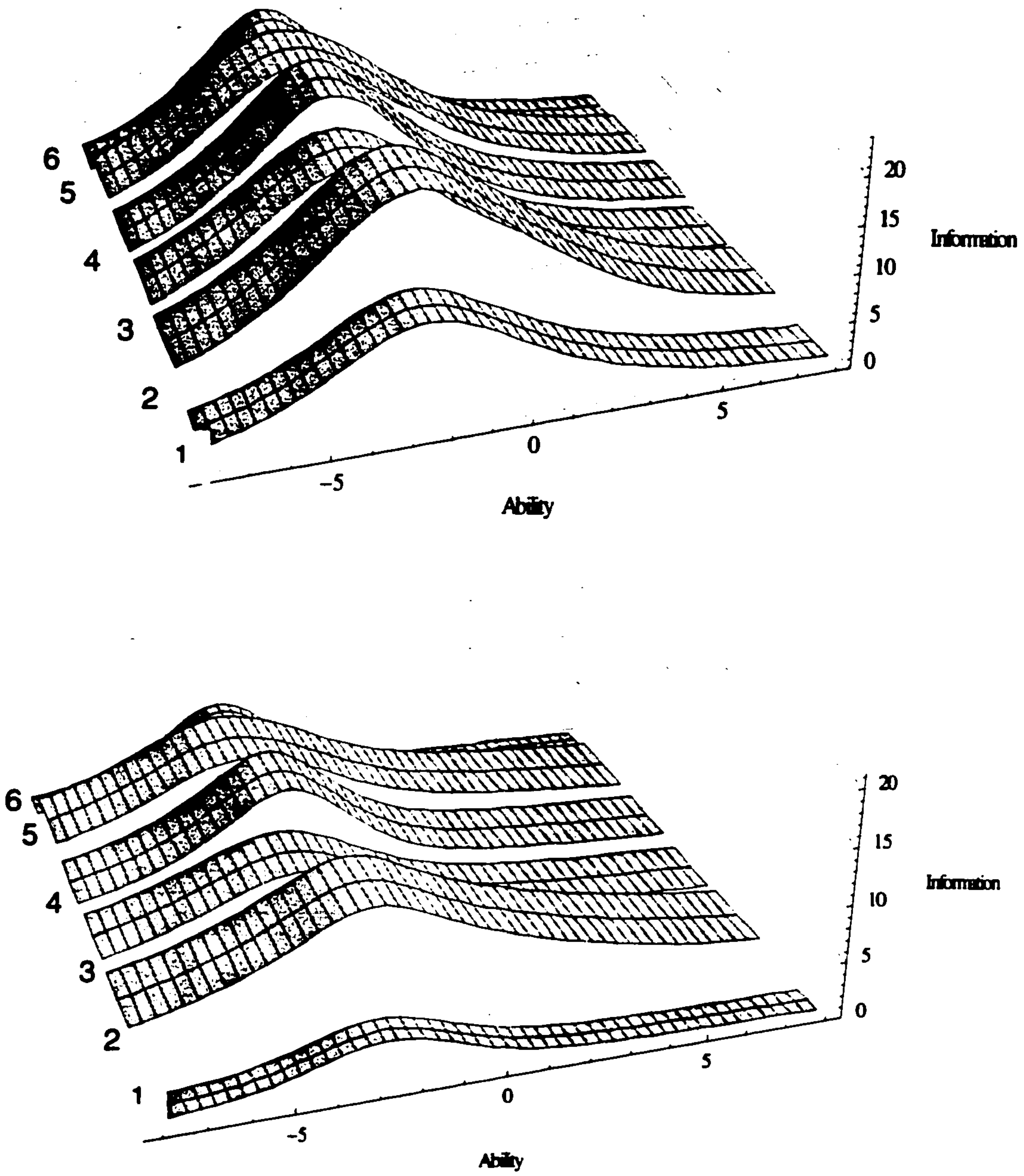


Figure 28. The MCC Licensure Examination Item Bank with 1,973 Items by Blueprint area.

References

Ackerman, T. (1989, March). An alternative methodology for creating parallel test forms using the IRT information function. Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco.

Ackerman, T. (1990, April). An evaluation of the multidimensional parallelism of the EAAP mathematics test. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Ackerman, T. (1991, April). An examination of the effect of multidimensionality on parallel forms construction. Paper presented at the annual meeting of the National Council for Measurement in Education, Chicago.

Ackerman, T. (1994). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education, 7, 255-278.

Adema, J. J. (1992). Methods and models for the construction of weakly parallel tests. Applied Psychological Measurement, 16, 53-63.

Adema, J. J., & van der Linden, W. J. (1989). Algorithms for computerized test construction using classical item parameters. Journal of Educational Statistics, 14, 279-290.

Alberta Learning. (1999). Achievement testing program. Edmonton: Alberta. Government of Alberta.

Armstrong, R. D., Jones, D. H., & Kunce, C. S. (1998). IRT test assembly using network-flow programming. Applied Psychological Measurement, 22, 237-247.

Armstrong, R. D., Jones, D. H., & Wang, Z. (1994). Automated parallel test construction using classical test theory. Journal of Educational Statistics, 19, 73-90.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Thousand Oaks: Sage.

Carbonaro, M. D. (1988). Computerized test item banking: Features. Unpublished master's thesis, University of Alberta, Edmonton, Canada.

Embretson, S. E. (1996). The new rules of measurement. Psychological Assessment, 8, 341-349.

Gibson, W. M., & Weiner, J. A. (1998). Generating random parallel test forms using CTT in computer-based environment. Journal of Educational Measurement, 35, 297-310.

Gierl, M. J. (1998). Item banking: Structural features and implementation issues. Paper presented at the annual meeting of the Canadian Society for the Study of Education, Ontario, Ottawa.

Gulliksen, H. (1950). Theory of mental test. New York: McGraw-Hill.

Hambleton, R. K. (1980). Contributions to criterion-referenced testing technology: An introduction. Applied Psychological Measurement, 4, 421-424.

Hambleton, R. K. (1986). The changing conception of measurement: A commentary. Applied Psychological Measurement, 10, 415-422.

Hambleton, R. K., & De Gruijter, D. N. (1983). Application of item response models to criterion-referenced test item selection. Journal of Educational Measurement, 20, 4, 355-367.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 3, 159-170.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Hingham, MA; Kluwer Nijhoff

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. NewBury Park, CA: Sage.

Kolen, M. J., & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer.

Lin, C., & Spray, J. (2000, April). Automated test assembly: Comparisons between classical test theory and item response theory in assembling parallel forms. Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 193-198.

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

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

Luecht, R. M. (1998a). CASTISEL 3PL [Computer Program and manual].

Luecht, R. M. (1998b). Computer-assisted test assembly using optimization heuristics. Applied Psychological Measurement, 22, 224-236

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. Journal of Educational Measurement, 35, 229-249.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ. Erlbaum.

Medical Council of Canada. (1999). Information pamphlet computer-based testing: Qualifying Examination. Available: www.mcc.ca/qeidescription.html.

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

Mislevy, R. J., & Bock, R. D. (1997). BILOG 3.11: Item analysis and test scoring with binary logistic test models [Computer Program]. Morrseville, IN: Scientific Software

Reckase, M. D. (1997). The past and future of multidimensional item response theory. Applied Psychological Measurement, 21, 25-36.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. Psychometrika, 42, 193-198.

Sanders, P. F., & Verschoor, A. J. (1998). Parallel test construction using classical item parameters. Applied Psychological Measurement, 22, 212-223.

Sands, W. A., & Waters, B. (1997). Introduction to ASVAB and CAT. In W. A. Sands, B. K. Waters, & J. R. McBride. (Eds.), Computerized adaptive testing: From inquiry to operation. Washington DC: American Psychological Association.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17, 277-292.

Stocking, M. L., Swanson, L., & Pearlman, M. (1993). Application of an automated item selection method to real data. Applied Psychological Measurement, 17, 167-176.

Theunissen, T. J. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Theunissen, T. J. (1986). Some applications of optimization algorithms in test design and adaptive testing. Applied Psychological Measurement, 10, 381-389.

Timminga, E. (1998). Solving infeasibility problems in computerized test assembly. Applied Psychological Measurement, 22, 280-291.

Timminga, E., & Adema, J. J. (1996). An interactive approach to modifying infeasibility 0-1 linear programming models for test construction. In G. Engelhard Jr. & M. Wilson (Eds.), Objective measurement: Theory into practice (Vol. 3). Norwood NJ: Abex.

van der Linden, W. J. (1986). The changing conception of measurement in education and psychology. Applied Psychological Measurement, 10, 325-332.

van der Linden, W. J. (1996). Assembling tests for the measurement of multiple traits. Applied Psychological Measurement, 20, 373-388.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. Applied Psychological Measurement, 22, 211.

van der Linden, W. J. (Ed.) (1998). Optimal test assembly [Special Issue]. Applied Psychological Measurement, 22 (3).

van der Linden, W. J. (2000). Optimal assembly of tests with item sets. Applied Psychological Measurement, 24, 225-240.

van der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. Journal of Educational Measurement, 35, 185-198.

van der Linden, W. J., & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Guliksen's matched random subtests method. Applied Psychological Method, 12, 201-209.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). Handbook of modern item response theory. New York: Springer.

van der Linden, W. J., & Luecht, R. M. (1998). Observed-Score equating as a test assembly problem. Psychometrika, 63, 401-418.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. Applied Psychological Measurement, 22, 259-270.

Veldkamp, B. P., & van der Linden, W. J. (2000). In Wim J. van der Linden & Cees A. W. Glas (Eds.) Computerized Adaptive Testing: Theory and Practice. Boston: Kluwer Academic Publishers.

Wainer, H. (1990). Introduction and History. In H. Wainer (Ed.), Computer Adaptive Testing: A Primer. (pp. 1-21). New Jersey: Lawrence Erlbaum.

Wightman, L. F. (1998). Practical issues in computerized test assembly. Applied Psychological Measurement, 22, 292-302.

Wright, B. D., & Bell, S. R. (1984). Item Banks: What, Why, How. Journal of Educational Measurement, 21, 331-345.

Wolfram Research (1999). Mathematica (Version 4.0) [Computer software] Champagne, Illinois: Wolfram Research.

Appendix A

The ratio method used in this dissertation is biased when integrating across the entire theta range. Specifically, this measure divides one curve over the other and then adds the inverse (i.e., A/B + B/A) across the theta continuum. For example, 0.1/0.2 + 0.2/0.1 equals 2.5 while 0.8/0.9 + 0.9/0.8 equals 2.0. Therefore, comparisons between curves across the entire theta range will be biased. This ratio measure is used to help determine differences between curves that are consequential. Thus, tests with higher values in information or true scores will produce small differences between the curves because the standard error of estimation will be minimal. Conversely, however, tests with lower amounts of information or true scores will have larger standard errors and thus any differences between test curves will be magnified. Therefore, this ratio measure, when applied across the full theta range, should be used in conjunction with the graphical methods described in this study in order to see where the differences occur across the theta continuum. Note that the ratio measure is less biased when the curve comparisons are 1 standard deviation away from the cut-point because the information or true score changes between the curves will be smaller.