AN ABSTRACT OF THE THESIS OF

Emily Lynne Cole for the degree of Master of Science in Movement Studies for the Disabled presented on August 11, 1989.

Title: An Application of Item Response Theory to the Test of Gross Motor Development

Abstract approved: Redacted for Privacy
(John Dunn, Ed.D.)

Abstract approved: Redacted for Privacy
(Terry Wood, Ph.D.)

The purposes of this study were to (a) provide insight into the use of item response theory (IRT) with psychomotor skills, (b) assess the psychometric properties of the Test of Gross Motor Development (TGMD) using IRT, and (c) provide a basis for future studies of the TGMD using IRT. The dichotomously scored TGMD is a test instrument which measures psychomotor skills in a framework similar to cognitive tests, thus providing a convenient "transitional" type test which can be used to examine the use of IRT with psychomotor skill tests. The present study employed data used by Ulrich (1985) in the original psychometric analysis of the TGMD. The data consisted of 913 subjects aged 3 to 10 years, nonhandicapped and 20 mildly mentally handicapped. Since IRT cannot provide accurate ability estimates at mastery levels of 0% and 100% mastery, 32 subjects were deleted from the record. Since the TGMD was found to be multidimensional, the test was analyzed by subtests so not to violate the unidimensionality assumption of IRT.

Interpretation of traditional item statistics using classical test theory (CTT) and IRT item parameters revealed that item difficulty and item discrimination were closely related. The locomotor IRT difficulty parameters revealed a high negative correlation (r = -.87) with the CTT difficulty statistics, while the object control IRT difficulty parameters displayed a very high negative correlation (r = -.98) with their CTT

counterparts. Item response theory discrimination parameters correlated highly with CTT discrimination statistics within the locomotor (r = .91) and the object control (r = .94) subtests.

IRT analysis revealed that the locomotor subtest was less difficult (median difficulty = -.944) than the object control subtest (median difficulty = .053) and the object control subtest displayed a better discrimination index (median = 2.17) than the locomotor subtest (median = 1.54). In addition to difficulty and discrimination indices, IRT also provided the amount of information given by each item and subtest, which indicated the precision in measuring various ability levels. The locomotor subtest information was reported at I = 15.50, indicating adequate precision to measure low ability (Θ = -1.857). The object control information function showed that the subtest displayed more information (I = 18.24) at a slightly higher ability level (Θ = -1.643).

The results of the item analysis revealed that all items (behavioral criteria) of the hop, leap, and the overhand throw displayed effective psychometric properties, while 9 out of 12 skills contained items that displayed poor psychometric characteristics and/or did not fit the two-parameter model. The run (items 1, 3, and 4), gallop (items 6 and 8), horizontal jump (item 18), skip (item 20), slide (items 23, 24, 25, 26), strike (items 27, 28, and 29), stationary bounce (item 32), catch (item 34 and 35), and the kick (item 38) should be revised.

Since the TGMD is also used as a criterion-referenced test the decision validity of the mastery classification cut-off scores was analyzed. For the purposes of these analyses true mastery state was determined by IRT because it provides an estimation of underlying ability. It was found that IRT and CTT showed a high agreement of classifying masters and nonmasters at the 70% and 85% levels of mastery. The locomotor subtest revealed a decision validity coefficient of .93 and .99 for the 70% and 85% mastery levels, respectively. The object control subtest revealed higher decision validity coefficients of .99 and .997 for the 70% and 85% mastery levels, respectively.

The TGMD subtests were found to best measure very low mastery levels, where the most precision for measuring ability represented 30% and 45% mastery for the locomotor and object control subtests, respectively.

Item response theory has been successfully employed in the cognitive and affective domains and shows great promise for the psychomotor domain. The present study set forth evidence that the IRT two-parameter logistic model provides an effective psychometric analysis of dichotomously scored psychomotor skills. The theory addresses many of the shortcomings of CTT, such as the inability to generalize item statistics to various populations and determine the contribution of test items independently. The invariance property of IRT is very appealing to those who must assess atypical populations because a single test can accommodate various populations and wide ranges of ability. The results of this study provide evidence that the characteristics of IRT are well suited to improve measurement and evaluation in the psychomotor domain.

An Application of Item Response Theory to the Test of Gross Motor Development

by

Emily Lynne Cole

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Completed August 11, 1989
Commencement June 1990

APPROVED:

Redacted for Privacy

Professor of Exercise and Sport Science, in charge of major (co-major professor)

Redacted for Privacy

Professor of Exercise and Sport Science, in charge of major (co-major professor)

Redacted for Privacy

Chair of Exercise and Sport Science

Redacted for Privacy

Dean of Health and Human Performance

Redacted for Privacy

Dean of Graduate School

Date thesis presented August 11, 1989

DEDICATION

This study is dedicated to my grandparents: Mr. John W. Cole, Mrs. Cecile Cole, Mr. James C. DeLange, and Mrs. Arlene S. DeLange for their foresight and wisdom that helped me achieve my goals.

AKNOWLEDGEMENTS

I would like to express my deep appreciation to the co-chairmen of my thesis committee, Dr. John Dunn and Dr. Terry Wood. Thanks to Dr. Dunn for his genuine interest in my professional growth, academic guidance, and expertise in issues regarding special populations. Without the efforts and expertise of Dr. Terry Wood this study would not have been possible. Through high expectations, encouragement to tackle new frontiers, and the support to make this project a success, Dr. Wood helped make this project a challenging and rewarding experience. I would also like to thank the other members of my committee, Dr. John Ringle and Mr. Ken Kosko for their efforts, suggestions, and encouragement.

I would like to express my sincere gratitude to Dr. Dale Ulrich for contributing the data for the present study and his sincere interest in the test analysis.

Special thanks are extended to Caron Shake, Diana Allen, Susan Hobble, LaJean Lawson, and members of the H-Club for their suggestions, friendship, and professional camaraderie.

I would like to express my sincere thanks to Steven Ovalle for his faith in my abilities, confidence in me to solve problems, and a special friendship that enhanced my potential to succeed throughout my studies at Oregon State University. Thanks Bud!

And finally, to my parents:

> "...for attaining wisdom and discipline;
> for understanding words of insight;
> for acquiring a disciplined and prudent life, doing what is right
> and just and fair;
> for giving prudence to the simple, knowledge and discretion to the
> young-

> Let the wise listen and add to their learning, let the discerning get guidance...
> Listen my child to your father's instruction and do not forsake your mother's teaching.
> They will be a garland to grace your head and a chain to adorn your neck" (John 1: 2-5, 8-9).

Thank you Mom and Dad for all your wisdom, enthusiasm to try new things and accomplish goals. But most importantly, all your love.

Thank you all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDIX FIGURES

An Application of Item Response Theory to the Test of Gross Motor Development

CHAPTER 1

INTRODUCTION

The assessment of motor skills provides information from which important educational decisions are made regarding programming and placement. Thus, it is important that standardized, valid, and reliable assessment tools be developed to provide appropriate physical education programs for all children. Since physical education services are desirable for all children, tests used for physical education purposes must also assess a wide range of abilities. Evaluations must be specific to the area of educational need and students suspected of having motoric difficulties must be assessed by knowledgeable individuals (e.g., physical educators or related personnel such as physical therapists).

With the current index of assessment tools, physical educators are challenged in their efforts to meet the requirements of good testing practices. Seaman (1988) and Baumgartner and Horvat (1988) have identified several major problems in assessing students with handicaps. Measuring motor behavior parameters of handicapped populations is difficult due to the limitations of current test instruments, characteristics of atypical populations, and lack of appropriate measurement and evaluation training for physical education professionals (Baumgartner & Horvat, 1988). Many tests are limited in their application, because they test only a defined range of age, ability, and handicapping condition. Norms for handicapped and nonhandicapped students on the same test are needed for placement decisions, but are not readily available (Baumgartner & Horvat, 1988). Scoring of motor skill tests are neither sensitive to measuring students of low ability, nor are they able to measure change in performance over time (Seaman, 1988). For administration purposes, it is easier to evaluate a student's ability based on one composite score. These scores, however, are not the most accurate estimation of one's ability. Composite scores also have been criticized for their lack of application to program development since specific problem areas cannot be detected (Davis, 1984; Klesius, 1981). In addition, teachers must often administer many tests for a complete assessment of one or a group of individuals. It is not uncommon for an adapted physical education teacher to administer a combination of fitness, motor development, and/or motor proficiency tests. Administering many tests is time consuming, expensive, and reduces instruction time.

Further problems in the measurement of psychomotor skills stem from the general lack of knowledge and understanding of assessment tools by those who administer them. Professionals have expressed dissatisfaction with using standardized tests (Cratty, 1986; Davis, 1984), replacing them with simple observation. Measurement tools are valid only for specific populations and purposes identified by the test developer. Tests are rendered invalid if assumptions inherent in the test guidelines are not met. In summary, physical educators find it difficult to validly and reliably assess handicapped students. It is likely that some important educational decisions are made based on inappropriate measurement and evaluation practices.

An ideal psychomotor test should be valid for virtually all school students and should serve either as a screening and/or a diagnostic tool. Tests need to accurately discriminate among ability levels. For the purpose of screening, a test should discriminate those needing special services from those who would benefit from placement in regular educational programs. Good tests provide normative comparisons enabling professionals to judge ability relative to the norm, and criterion-referenced standards which enable comparisons to a predetermined performance. Criterion-referenced tests also provide information that could be incorporated directly into programming. Additionally, test developers should consider cost, amount of special equipment, and administration time. To gain notoriety among teachers and professionals, a test must be relatively easy to administer. Directions and scoring methods should be clearly stated, and evidence of the test's validity and reliability should be stated in the test manual.

The Test of Gross Motor Development (TGMD) designed by Ulrich (1985) is a recent, well-designed addition to the repertoire of tests in the psychomotor domain. The TGMD is a popular screening test which assesses gross motor ability of children from 3 to 10 years of age. The test contains many attributes of a well-designed assessment tool. It is economical, requires no special equipment, and is relatively quick to administer. The TGMD is both a norm-referenced and a criterion-referenced test. The normative data may be used for making placement decisions, while the criterion-referenced data may be utilized in individual education plans and class activities. Langendorfer (1986) contends that the TGMD represents a new significant addition to the current index of psychomotor tests.

One of the most common shortcomings of psychomotor tests is the lack of generalizability to assess broad ranges of ability. A test is not useful if norms are not appropriate or the behavioral criteria insensitive to ability levels. These problems may be inherent in the test theory used to develop the assessment tool.

Most psychomotor test development is based on classical test theory (CTT) which focuses on the interpretation of test scores. The interpretation of test scores depends on the frame of reference. Typically norm-referenced or criterion-referenced measures are used in physical education. Norm-referenced tests estimate an examinee's ability in relation to the performance of the referenced group. Criterion-referenced tests estimate an examinee's ability based on the amount of the specified domain mastered. Referenced scores are based on the performances of the standardized sample. Thus, with CTT, test parameters may vary depending on the group tested. Because the veracity of test parameters is limited to the defined population and specific use of the test, physical educators who assess atypical populations find it difficult to use the tests currently available.

The contribution of individual test items to the estimation of underlying ability cannot be determined using CTT. This creates problems for those who must omit test items when testing students with physical limitations. Such omissions may render a test invalid. The estimation of ability depends on the selection of test items, but the individual contribution of test items cannot be determined. Tests must be revalidated when the pool of test items is changed. Tests based on CTT are limited to specific populations and intended uses. A major limitation of CTT is the inability to analyze test items independently.

Item response theory (IRT) is an alternative test theory that shows great promise for application to the psychomotor domain (Costa, 1986; Disch, 1987; Safrit, 1987; Spray, 1987; Wood, 1987) and may help to resolve many of the current problems in the measurement and evaluation of motor skills. Item response theory focuses on each test item, rather than the test score. It is a mathematical model that relates item probability of success to one's ability level (Hambleton & Cook, 1977), producing a measurement tool that is invariant across tests and subject populations (Lord, 1980). Item response theory provides information on the individual's underlying ability (e.g., gross motor development) relative to the item characteristics, whereas the more traditional approach, CTT, estimates an individual's ability in comparison to the performance of others. Item response theory does not limit measurement to the original referenced group. Item response theory addresses the population limitations of CTT and the inability of CTT to analyze the individual contribution of test items.

Item response theory has had a dramatic effect on measurement in the cognitive and affective domains (Loyd, 1988; Marco, 1977). It has been used primarily for written test construction, but may be employed for item analysis, pattern analysis, item banking, adaptive testing, test equating, redesigning tests, optimal score weighting, and relating developmental age to chronological age. Item response theory may also be used to enhance classical forms of test validity (Wood, 1987).

To further examine the potential of IRT models for improving measurement and evaluation in the psychomotor domain, Wood (1987) recommended research to include: (a) research papers and discussion of the applicability of IRT to improve tests in the psychomotor domain; (b) examination of IRT literature in the affective and cognitive domains accompanied by analysis of such tests in physical education using IRT methodology; (c) application of current models and/or the development of new IRT models specific to the psychomotor domain; (d) investigation of IRT models and the robustness of the assumptions; and (e) research concerning the practical application of IRT models to motor behavior measurement tools. Although the foremost application of IRT has been in education and psychology, recent discussions and employment of IRT models to the psychomotor domain offer promise for improved measurement and evaluation in physical education.

Item response theory was the focus in a recent issue of Research Quarterly for Exercise and Sport (September, 1987). Spray (1987) reviewed IRT and its potential use in physical education. Disch (1987), Safrit (1987), and Wood (1987) discussed the advantages, disadvantages, and potential problems in meeting the assumptions of IRT with psychomotor test data. Review of these papers suggests that IRT is a powerful test theory that warrants empirical research to determine its applicability to the psychomotor domain.

Safrit (1987) reported two studies analyzing bowling skill data with IRT models (Costa, Safrit, & Cohen, 1987; Safrit, Costa, & Cohen, 1987 cited in Safrit, 1987). The results of these studies are encouraging since Safrit speculates that IRT may be applied to other psychomotor skills. Steffens, Semmes, Werder, & Bruininks (1987) used the Rasch scaling model of IRT to equate quantitative and qualitative measures of motor development. The IRT data were used to equate scores of the Bruininks-Oseretsky Test of Motor Proficiency (Bruininks, 1978) and The Motor Skills Inventory (Werder & Bruininks, 1988) to arrange scores along a common scale of difficulty. This enabled the authors to combine qualitative and quantitative motor development data. The ability to determine the contribution of items to a test and invariance across populations makes IRT very powerful and may improve measurement and evaluation in the psychomotor domain.

Although the application of IRT to psychomotor skills is relatively recent, it may provide a revolutionary approach to measurement in physical education. Application of IRT to the TGMD will provide further opportunity to ascertain the usefulness of applying IRT models to motor skill tests. This study will provide a basis from which future studies may be pursued. Test parameters may be recalibrated with atypical populations to develop norms based on ability levels, employ test equating procedures, and pursue the potential of adaptive testing using motor skills. The present study follows a prescribed line of research which assesses the potential of IRT to improve measurement and evaluation of motor behavior.

Significance of the Study

This study will provide insight to the usefulness of the application of item response theory to psychomotor skills. Item response theory will be applied to assess the usefulness of the Test of Gross Motor Development for subjects of varying ability. This study may provide an opportunity for future investigations. For example, studies might be designed to:

- Validate the composite score through the use of weighted items
- Employ pattern analysis to examine strong and/or weak areas of motor development
- Study the utility of comparing TGMD scores across handicapped populations

- Develop norms based on ability levels
- Equate developmental age versus chronological age using the TGMD
- Examine alternative criteria to provide a database of items for various ability levels
- Employ test equating techniques to compare the TGMD to other test batteries.

Statement of the Problem

The purpose of this study was to examine the psychometric properties of the Test of Gross Motor Development through the use of IRT. Item response theory was employed to determine item discrimination, item difficulty, the precision of test items at various ability levels, and to assess the accuracy of criterion-referenced cut-off scores.

Research Questions

The TGMD was analyzed by the one-parameter and two-parameter logistic item response models to:

(a) Assess the fit of the IRT models to the TGMD data;

(b) Determine difficulty and discrimination parameters of test items and subtests;

(c) Estimate the degree of precision with which the subtests estimate true ability using standard error of estimation;

(d) Estimate the item and test information curves to determine the ability level at which a given item and subtest provides the most information;

(e) Validate the 70% and 85% mastery classification cut-off score of the subtests.

Definitions

Gross Motor Development. Gross motor development is the gradual increase in temporal and spatial coordination of the total body and larger muscle groups in simultaneous movement (Williams, 1983). Gross motor skills are the basis for locomotion and object manipulation skills.

Item. An item will be defined as a single behavioral criteria of a skill. Three or four items comprise a motor skill on the TGMD.

Subtests. There are two subtests on the TGMD: (a) the locomotor subtest which includes the run, gallop, hop, leap, horizontal jump, skip, and slide and (b) the object control subtest which includes the two-handed strike, stationary bounce, catch, kick, and overhand throw skills.

Gross Motor Development Quotient (GMDQ). A standard score that indicates general motor ability. It is a composite score of the 12 skills.

Assumptions

The following assumptions will be made regarding subjects, motor development tests, and IRT. It is assumed that motor development is a sequential progression of gross motor skills, with time of skill acquisition as a variable. Motor development tests also assume that the behavioral criteria or test items are biomechanically correct for all children. The subjects' performance on the TGMD is expected to be representative of their underlying gross motor ability.

CHAPTER 2

REVIEW OF LITERATURE

Introduction

The purpose of this chapter is to review the literature on current assessment practices in adapted physical education from a practical and theoretical view. First, a discussion of the challenge in meeting the legal mandates of assessing children with handicaps with the current index of test instruments will be reviewed. Second, the TGMD will be presented, since it will serve as the dependent variable in this study. Third, classical test theory and item response theory will be contrasted. Problems of assessment may be the function of the test theory that serves as the foundation of the test. A discussion of the advantages and disadvantages of classical test theory and item response theory will be given. Lastly, the application of IRT to the psychomotor domain, more specifically as it applies to The Test of Gross Motor Development, will be examined.

Current Assessment Practices

The Education for All Handicapped Children Act of 1975, PL 94-142 (EHA) has enhanced the need for motor skill assessments which, in turn, has had a major impact on the subsequent development and implementation of psychomotor tests. The EHA extends the right of a free and appropriate public education to children regardless of severity of condition or handicap. Physical education is specifically cited in EHA's definition of special education as a mandated curriculum area. All handicapped children must be provided physical education services. Since the EHA has been enacted, it is estimated that 80% of all handicapped children are served in regular schools and classrooms (Baumgartner & Horvat, 1988). This increases the number of students that must be screened for placement and tested for diagnostic purposes. The

EHA further addressed the issue of test characteristics and testers. Under the regulations of the EHA, tests must be standardized, valid, and reliable for their intended populations. Evaluations must be specific to the area of educational need and students suspected of having motor difficulties must be assessed by knowledgeable individuals (e.g., physical educators or related personnel such as physical therapists). The EHA also requires test conditions to elicit a subject's best ability or aptitude. It is important that valid and reliable assessment tools be developed to provide appropriate physical education programs for all children.

The EHA mandates that students be taught in the least restrictive environment. The least restrictive environment is the setting in which the student can best learn. Screening and diagnostic tests provide objective measures that may be used to determine the least restrictive environment. Screening serves to initially identify children who have difficulties participating in the regular physical education curriculum and may qualify for special physical education services. Screening tests may be routinely given to children entering school or used to determine if there is probable cause for further diagnostic testing. Diagnostic tests tend to be comprehensive, test a variety of potential disability areas, and aim to identify the area(s) of special need. Information from assessments is used to make placement and programming decisions.

Ideally, psychomotor tests would be valid for virtually all school students and should serve either as screening and/or as diagnostic tools. To assess all ranges of ability, tests need to discriminate well among ability levels. Good tests should provide both normative comparisons enabling professionals to judge ability relative to the norm, and criterion-referenced measures which enable comparisons with the criterion performance.

Although no test is perfect, some exhibit good test construction and relative ease of administration. The test developer should consider cost, amount of special equipment, and administration time. To gain notoriety among teachers and professionals in the field, a test must be relatively easy to administer. In addition, directions and manner of scoring should be clearly stated. Evidence of the test's validity and reliability should be stated in the test manual.

With the current index of assessment tools, physical educators are faced with a challenge in meeting the requirements of the EHA (Seaman, 1988). The EHA demands well-constructed tests to effectively measure students in the public school system who are suspected of having a disability. But, "faced with a 'mixed bag' of available motor tests, the physical educator may find test selection a frustrating dilemma" (Werder & Kalakian, 1985, p. 31). Seaman (1988) and Baumgartner and Horvat (1988) discussed problems in meeting the assessment demands of the EHA. Dilemmas arise due to the limitations of current tests, difficulties in measuring handicapped populations, and the limited knowledge of professionals regarding measurement and evaluation practices in physical education (Baumgartner & Horvat, 1988).

Many tests are limited in their application, encompassing only a limited age, ability range, and handicapping condition. Physical education teachers must assess a wide range of student abilities, but are limited by a shortage of well-normed motor ability test instruments. Tests with norms for handicapped and nonhandicapped students are needed for placement decisions, but are not readily available (Baumgartner & Horvat, 1988).

Scoring of many motor skill tests poses a problem, in that the tests are neither sensitive in measuring students of low ability, nor are they able to measure change in performance over time (Seaman, 1988). Composite scores are commonly used by test developers and practitioners to estimate students' ability, but these scores are not the most accurate representation of ability (Davis, 1984; Klesius, 1981). For administration purposes, it is easier to evaluate a student's ability based on one composite score. Unfortunately, composite scores do not provide the most information regarding one's ability. Composite scores may mask abnormal scores which may lead to inappropriate placement decisions. A student may have an abnormally high score on one subtest and an abnormally low score on another subtest, but his or her composite score may be in the normal range. Composite scores also have been criticized for their lack of application to program development since specific problem areas cannot be detected (Davis, 1984; Klesius, 1981).

Measuring motor skills of populations with handicapping conditions is difficult due to the large variability of abilities within and between special populations. Applying traditional methods of test development to handicapped populations is tedious and cumbersome because tests must be validated for each population within defined ability ranges. Consequently, adapted physical educators are required to use a variety of tests that match the population and ability of their students. However, tests do not typically provide norms for all handicapped populations found in the public schools and teachers must use tests that do not provide valid interpretations of their students' ability. A common practice is to administer tests that were validated on nonhandicapped populations and make inferences regarding a student's motor ability relative to the nonhandicapped norm. These scores may be useful in determining how a student will perform in the mainstream setting, but they do not provide information that can be useful in developing programs. Few tests provide a general indicator of how well a student will perform in the mainstream setting, thus, teachers must administer many tests for a complete assessment of one or a number of individuals.

Further problems in measurement of psychomotor skills lie in the lack of general knowledge and understanding of assessment tools by those who administer them (Baumgartner & Horvat, 1988; Davis, 1984; Klesius, 1981; Seaman, 1988). Professionals often lack the knowledge to critique assessment tools, use tests appropriately, and accurately interpret the results. In addition to the lack of knowledge of practitioners, information regarding assessment of handicapped populations is not readily available. In summary, physical educators find it difficult to meet the regulations of the EHA with the current index of psychomotor tests. Consequently, inappropriate measurement and evaluation practices may be influencing important education decisions.

The Test of Gross Motor Development

The TGMD was designed to meet the assessment demands of the EHA and practitioners. The TGMD is an individually administered test that evaluates gross motor functioning of children from 3 to 10 years of age. The TGMD consists of 12 gross motor skills commonly taught in preschool and elementary physical education classes. The skills are divided into two subtests: locomotor and object control. The locomotor subtest includes the run, hop, leap, gallop, horizontal jump, skip, and the slide. The object control subtest includes the two-handed strike, stationary bounce, catch, kick, and the overhand throw. Each skill is measured by 3 or 4 behavioral criteria which will be referred to as items. Items comprising each skill represent mature motor patterns. Higher scores reflect more mature gross motor development.

Ulrich (1985) proposed five uses for the TGMD. He projected that it may be used (a) as a screening tool to detect possible motor deficiencies, (b) to provide instructional and prescriptive programming for individualized education plans, (c) to evaluate student progress, (d) to assess motor skill program effectiveness, and (e) as a research tool.

Items are dichotomously scored, pass/fail (1 or 0). Three to four performance criteria comprise a skill. An examinee receives a raw score for the total number of criteria exhibited. Since the number and difficulty of the criteria vary between tasks, raw scores are not comparable. Raw scores are used to locate standard scores. The

TGMD provides standard score norms for each subtest and the composite score (the gross motor quotient). Percentile rank scores are useful to compare a student's performance between subtests and other tests. The Gross Motor Development Quotient (GMDQ) is a composite score of the sum subtest standard scores which is converted to a quotient utilizing tabled values.

Validation of the TGMD

Content-related evidence of validity was established using Safrit's (1981) protocol for determining logical validity. Logical validity assesses the "...the degree to which the components of skills measured by the test correspond with those identified by the test developer" (Safrit, 1981, p. 55). The reviewer compares the test developer's intended purpose and the defined skill components to those actually measured by the test. If the test measures what it claims to measure, the test has logical validity. Once logical validity has been assessed, the test is evaluated for its content relevance. Content relevance evaluates components of a motor skill test to determine if important content is omitted, unimportant content is included, and determines if items are appropriately weighted. Ulrich sought the opinion of three content experts to judge the TGMD's logical validity. The experts unanimously agreed that the skills accurately represent the gross motor domain.

Construct-related evidence is the degree to which the test measures the construct, and/or underlying traits it is designed to measure (Wood, 1989). Fundamental gross motor development is the construct measured by the TGMD. Ulrich provided evidence for construct validity by testing three hypotheses: (a) "...that the principle underlying structure of the test would reflect gross motor development..", (b) "...that gross motor development would improve significantly across age levels...", and (c) "...mentally retarded children would score lower on the TGMD..." (Ulrich, 1985, p. 31).

A factor analysis study was performed to test the first hypothesis that the test's principle underlying factor was gross motor development (Ulrich, 1985). "Factor analysis is a multivariate statistical technique that identifies common sources of variation among a set of theoretically related variables. The goal of factor analysis is to obtain a parsimonious number of factors that can be logically related to the variables" (Disch, 1989, p. 159). Ulrich used a principal components analysis with a varimax rotation, where loadings lower than .4 were deleted from the results. Three dimensions explaining 75% of the common variance were found. Utility of the three-factor structure was examined using a percent of variance criterion of 70% and the scree test. Nine skills, from both subtests (hop, leap, jump, slide, strike, bounce, catch, kick, and throw) loaded on the first dimension, accounting for 62% of the total common variance. The factor loadings for these skills ranged from .59 to .83. Six locomotor skills (gallop, hop, leap, jump, skip, and slide) loaded on the second factor, accompanying 8% of the total common variance and factor loadings ranging from .44 to .78. The run task loaded on a third dimension with 5% of the total common variance and a factor loading of .89. Ulrich concluded that the skills are highly related because 75% of the common variance shared by all 12 skills is explained by these factors.

Ulrich intended the TGMD to measure one construct, gross motor development, through the use of two subtests; object control and locomotor. If this were so, then one would expect the factors to have a simple structure by loading on two primary factors (locomotor and object control domains). The factor analysis revealed that the skills load on three dimensions with four skills loading on the first two factors and the run loading on a third factor. The run could load on a separate factor because (a) it is a lower level motor program (Magill, 1985), thus not a good measure of motor development, or (b) it is the most unreliable skill on the test. The results from the factor analysis are confusing and do not lend support to the construct validity of the test.

Since gross motor development tends to increase with age, Ulrich used correlation and analysis of variance procedures to evaluate the second hypothesis. Subtest scores and composite raw scores correlated with chronological age revealing a close relationship between test performance and chronological age. An ANOVA using the same data revealed a statistically significant age effect ($p < .01$). Ulrich concluded that the TGMD measures gross motor development because the scores tend to increase with age.

Multivariate analysis of variance (MANOVA) yielded a significant group effect where a nonhandicapped population consistently scored higher on the TGMD than a handicapped population (Ulrich, 1986). Since performance on the TGMD increases with age and mental ability, Ulrich postulated that the TGMD measures gross motor development as its only construct.

The TGMD's norm-referenced stability and inter-scorer reliability were determined by a single generalizability study (Ulrich & Wise, 1984). Generalizability theory allows the investigator to analyze the relative contribution of specific sources of measurement error, or unreliability, with one study (Morrow, 1989). Potential sources of measurement error, called facets, are identified, and then tested through analysis of variance procedures. The relative variance of each facet is compared to the total variance which provides an indication of the unreliability for that facet (e.g., rater or occasions). Generalizability studies also provide generalizability coefficients which are an indication of the estimated reliability for projected universes of interest (Morrow, 1989). Ulrich and Wise (1984) employed a two facet design with rater (n=20) and testing occasions (n=2) acting as the generalization facets, while subjects (n=10) served as the facet of differentiation.

Test stability was measured by a test-retest procedure with a one-week interval between occasions. The estimated variance for the locomotor and object control subtests were .53 and .49, respectively. The percentage of variance associated with these subtests was an acceptable 1% (locomotor) and 2% (object control). The mean generalizability coefficients for the locomotor subtest were reported as .96, while .97 was reported for the object control subtest (see Table 1 for the ranges of generalizability coefficients).

Inter-scorer reliability was measured by the same generalizability study, with raters as the facet of interest. The estimated variance components for the locomotor subtest and object control were .53 and .81, respectively, while the percent of total variance was 1% for the locomotor subtest and 4% for the object control subtest. The mean generalizability coefficients for 2, 10, and 20 raters were reported as .86, .94, and .95, respectively.

Ulrich and Wise (1984) argued that the results of this study implied that the variability due to raters and occasions was relatively small compared to the variance for skill and composite scores across subjects. Although the mean generalizability coefficients for these two studies appear rather high, they should be interpreted with caution since the sample size was small (n=10).

Internal consistency was reported in the test manual for each subtest at each of the eight age levels. The subjects were selected from each of the eight age groups of the standardized sample. Ulrich (1985) used a split-half reliability procedure (n=25 per age level) yielding coefficients, adjusted with the Spearman-Brown prophecy formula, of .85 for the locomotor subtest and .78 for the object control subtest. All coefficients for each age group were statistically significant indicating that subjects of all ages tended to perform consistently on each item of the test.

The SEM was calculated for the subtest raw scores, subtest standard scores, and the composite standard score by using the formula $\text{SEM} = \text{SD}((1-r)^{-1/2})$; where the SD represents a standard deviation value and *r* is a reliability coefficient. The SEM for the subtest raw scores was calculated for each of the eight age groups with the standard deviation value computed from the standardized sample and the mean estimated test-retest generalizability coefficients from Ulrich and Wise (1984). The SEM for subtest raw scores at each age level were published in the test manual. The SEM for the subtest and composite standard scores were computed with standard deviations of 3 and 15, respectively, and the mean estimated test-retest generalizability coefficients from the subtest and composite scores. The SEM for the locomotor subtest was reported as .60 and the object control subtest SEM was reported as .52. The reported SEM for the composite score was .3.

Stability, interscorer reliability, internal consistency, and the standard errors of measurement (SEM) are summarized in Table 1.

Table 1

Norm-Referenced Reliability of the TGMD

		Locomotor		Object Control	
	Sample size	Mean	Range	Mean	Range
Stability	10	.96	.84-.99	.97	.95-.99
Interscorer	20	.95	.84-.99	.97	.96-.99
	10	.94	.79-.98	.96	.93-.98
	2	.86	.77-.93	.87	.80-.94
Internal consistency	25	.86	.83-.90	.78	.67-.93
SEM		.60		.52	

Composite score SEM = 3.0, SD = 15

Criterion-referenced reliability of the TGMD was determined by assessing the precision of the selected mastery cut-off scores. Ulrich (1984) established the reliability of mastery decisions in two studies: the first one examined the stability of classifying handicapped and nonhandicapped children; the second study assessed the stability of classifying preschoolers at cut-off scores that represented 45%, 50% and 60% mastery of the test items. A summary of results is presented in Table 2.

Based on the kappa statistics, it appears that the 70% mastery level is the most reliable for classifying nonhandicapped and moderately mentally retarded children over two occasions. The 60% mastery level was the most consistent for classifying preschoolers over two occasions. The results indicated that the composite score of the TGMD consistently classified masters and nonmasters in the gross motor development domain.

Table 2

Reliability of Mastery Classification Decisions

	Mastery Level	P	kappa
NH	85%	.89	.78
	70%	.92	.84
H	85%	.87	.62
	75%	.93	.83
Pre	45%	.92	.83
	50%	.91	.83
	60%	.89	.69

NH: nonhandicapped children, *n*=80
H: children with moderate mental retardation, *n*=40
Pre: preschoolers, *n*=53

Davis (1984) reported many professionals were not satisfied with the current index of assessment tools. He found that tests do not fit program needs, are restrictive to defined populations, and scoring systems are insensitive in measuring ability and change in ability. In contrast, the TGMD contains many components of a good test. Ulrich designed the TGMD in response to programming needs of physical education teachers; the 12 motor skills represent those most commonly taught in an elementary physical education setting. The TGMD's scoring system allows comparison of the composite score, with age, and the age at which the sample achieved 60% and 80% of the specific items for each skill. Ulrich (1984) also showed that the test discriminated between nonhandicapped children and children with mental retardation.

Klesius (1981) described good tests as standardized, economical, efficient in administration time, requiring little or no specialized equipment, and possessing the capacity to discriminate among ability levels. Ulrich's test meets these criteria. Davis (1984) and Klesius (1981) suggested that tests should measure characteristics independently to better assess areas of strength and weakness in contrast to composite scores which may mask a problem area. Ulrich provided norms for the composite score, percentile scores for object control and locomotor subtests, and the ages at which 60% and 80% of the sample mastered each item. The TGMD allows an examinee's composite score and performance on test items to be compared to the standardized sample data.

Although the TGMD appears to be well constructed, Langendorfer (1986) criticized the test's construct validity on the basis of Ulrich's omission of the term 'change' within the definition of gross motor development, and the subsequent omission of items that would reflect a change in gross motor development. In the TGMD manual, Ulrich cites Williams' (1983) definition of gross motor development as "...skillful use of the total body in large muscle activities that require temporal and spatial coordination of movement of a number of body segments simultaneously..." (p. 10). Langendorfer contends that Ulrich used Williams' definition of gross motor control rather than gross motor development.

Langendorfer also criticized the test items on the basis that they do not represent a progressive developmental sequence from immature to mature motor patterns, but rather describe mature skill performance. Theoretically, items could load on a gross motor control factor rather than gross motor development. Since gross motor control increases with age, and individuals with mental retardation could have significantly lower motor control, it could be argued that Ulrich's evidence for gross motor development may also support the construct of gross motor control. Langendorfer postulated the TGMD measures gross motor control because Ulrich fails to recognize 'change' as an integral component of the test's construct. Additional research is needed to examine Langendorfer's claims.

Ulrich argued gross motor development is estimated by the number of items mastered. He ascertained the TGMD accounts for change in performance since the test correlates with age, indicating test scores reflect gross motor development. Although the manual provides the age at which 60% and 80% of the standardized sample mastered test items (Ulrich, 1985, p. 17), Ulrich does not suggest the items provide a measure of gross motor development. The determination of motor development could be made in reference to the test items rather than the number of items attained. Item analysis may determine if a developmental trend is exhibited by the items.

Classical Test Theory

The shortcomings of psychomotor tests may be inherent not only in the design of the test, but also in the measurement theory which serves as the test's foundation. Current motor skill tests are based on traditional models of measurement, the most prevalent being classical test theory (CTT). Traditionally, CTT provides a framework from which "...test reliability, validity, and generalizability coefficients, standard errors of measurement and estimation, predicted true scores, and coefficients of attenuation" are computed (Spray, 1987, p. 204). Classical test theory is based on population parameters to determine an examinee's ability from an observed test score. The major focus is on the interpretation of test scores (i.e., what we infer from test data is important) and test validation is based on inferences made from an examinee's test score.

Classical test theory is based on weak assumptions underlying the nature of observed test scores and is useful when stronger assumptions cannot be met. The fundamental definition of CTT holds that an individual's test score, X, is a variable of the sum of one's true score, T, and the error of measurement, E, where $X = T + E$ (Safrit, 1981). The true score is an expected value if one was to repeat the test independently an infinite number of times. Through examination of standard error of measurement, one may determine the range in which the true score lies. It is assumed that: (a) the errors of measurement are unbiased, where $(E) = 0$, (b) the error and the true score are not correlated in any group, and (c) error is not related to the test items (Lord, 1980). In addition, errors of measurement are assumed to be equal for all ability levels (Hambleton, 1989). For example, low ability levels are estimated with the same magnitude of measurement error as the medium or high ability categories. Tests based on CTT often violate the constant error assumption. The magnitude of measurement error varies with items and ability. Since tests are typically designed to measure average ability, those in the lower or higher ability groups are tested with items that possess the most error and do not estimate ability efficiently. Because error is equally distributed across all items, it is not possible, with CTT, to determine error of measurement of individual test items or ability levels. Moreover, error is computed for the entire test and for a given population. Thus interpretation of test scores is limited to a defined population. Classical test theory relates an examinee's test score to scores obtained from the standardized sample, as seen in measures of central tendency, variance, and covariance (Lord, 1980). A test score is interpreted in relation to the performance of a group. Therefore, ability inferred from test scores varies with the group tested.

Classical test theory reliability and validity coefficients must be made in reference to the test score, rather than to individual test items. Thus, the contribution of each item on the test cannot be determined independently of other test items (Lord, 1977). If a student is physically impaired and cannot complete an item on the test, the test score is invalid. In contrast item response theory can estimate ability utilizing one item, but requires additional items for an accurate estimation. Classical test theory is limited in its application to specific populations, intended uses, and ability to analyze test items independently, which pose problems for practitioners who must assess atypical populations.

Item Response Theory

Item response theory (IRT) is an alternative test theory that can be used to enhance CTT. In contrast to CTT, which utilizes an individual's test score to make inferences about one's underlying ability, IRT utilizes individual test items as the basic unit of analysis. Item response theory provides a mathematical relationship between the probability of answering an item successfully and an examinee's ability level (Θ), allowing test data to be interpreted as an underlying trait or ability. Additional restrictive assumptions are made which allow test items and tests to be employed over a wide spectrum of examinees. Where CTT is based on weak assumptions regarding test scores IRT is based on strong assumptions (Spray, 1987). Under IRT, values for item parameters and an examinee's ability are invariant. With CTT, tests are not invariant across groups resulting in items that may be easy for one group and difficult for another. Item response theory estimates item difficulty and item discrimination parameters that are constant across populations. The major advantages of IRT are (a) items may be independently analyzed, and (b) tests are not population bound. These characteristics of IRT hold great promise for assessing atypical populations accompanying large ranges of ability.

Possible uses of IRT include item banking, adaptive testing (Spray, 1987; Spray, 1989), test equating (Skaggs & Lissitz, 1986), and assessing mastery cut-off scores (Spray, 1989). Since item parameters are known and invariant, items may be stored in item banks. Depending on the nature of the test, a test developer may choose items to assess specific ability levels. Tests may be designed (adapted) for each examinee by presenting items near an examinee's ability. Adaptive testing is efficient since the examinee is not given items which are too difficult or easy. This avenue of testing seems promising for assessing children with handicapping conditions. Item response theory has also been used to equate tests. Item and tests curves can be compared to determine if items and/or tests are similar in difficulty and discrimination. It is an effective method for test equating because test items are compared rather than the aggregate test score. Item and test parameters can also be compared across cultural groups to evaluate cultural bias (Cole & Moss, 1989). Test equating could also be used to measure bias in testing or identifying handicapped populations which may be helpful in making placement decisions. Item response theory has been employed in the development of criterion-referenced tests (Hambleton & deGruijter, 1983; Lord, 1980; Shannon & Cliver, 1987) and setting standard achievement levels (Kane, 1987). The test developer may choose items that have the most precision to measure abilities near the mastery cut-off score. These are just a few application areas of IRT; many more are cited in the literature.

Although IRT appears to address many shortcomings of CTT, its methodology is conceptually and mathematically rigorous (Wood, 1987). Item response theory is a complex process of estimating item and ability parameters from mathematical models which produce scales that have no direct relationship to the observed data (Loyd, 1988). In addition, IRT is based on the probability of success, which produces a scale ranging from 0 to 1, and ability (Θ) which is a standard score that ranges from $-\infty$ to $+\infty$. These measures are difficult to interpret for those not familiar with IRT. Lastly, IRT research is expensive and cumbersome since it requires a large number of subjects to generate stable item and test characteristic curves (Hambleton, 1989).

Although computationally rigorous, IRT has been successful in the cognitive and affective domains and more recently in the psychomotor domain (Charoenruk, 1989; Costa, 1986; Safrit, 1987; Steffens, Semmes, Werder, & Bruininks, 1988). However, the degree to which it may be applied to motor skills is not yet known.

Item response theory has four distinguishing characteristics and is based on two fundamental assumptions. The characteristics of IRT discussed by Hambleton and Swaminathan (1985) include:

(a) performance may be predicted or explained in reference to the proposed underlying ability or trait of the test;

(b) models of IRT specify a relationship between observable performance on an item and traits or abilities assumed to underlie each item or test;

(c) models provide a means to estimate scores for the underlying traits;

(d) traits are estimated by observable performance (or responses) on test items to provide a reference to underlying abilities.

The two fundamental assumptions underlying IRT are (Hambleton & Cook, 1977):

(a) Dimensionality of latent space depends on the number of traits that underlie test performance. Typically, models are unidimensional (items measure one trait), although multidimensional models have been discussed in the literature (Skaggs & Lissitz, 1986). Violations of unidimensionality ought to be considered in terms of item construct and difficulty (Loyd, 1988).

(b) Local independence may be defined by two forms, strong and weak. Classical test theory is based on weak assumptions where the components of a test score are uncorrelated. Item response theory is based on strong assumptions because an individual's performance is statistically independent of items on the test. Performance on an item must not effect performance on the other items of the test.

These "strong assumptions" may be difficult to meet. Moreover, the degree to which assumptions may be violated is not widely known. Research suggests that IRT models are fairly robust to violations of assumptions for knowledge and cognitive tests, but the conclusions put forth are not definitive. Research regarding the psychomotor domain is very limited and the degree to which assumptions can be met or violated in this domain is an avenue of future research.

The Item Characteristic Curve (ICC)

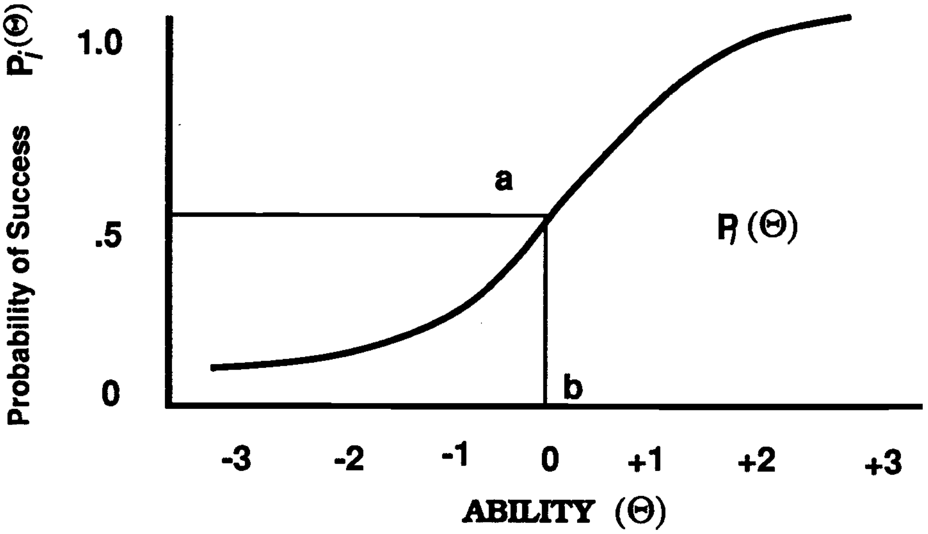
The ICC is a mathematical function that relates the probability of success (correctly responding to an item) to the underlying ability or latent trait measured by that item. It acts as the vehicle by which item and test parameters are interpreted. A curve is generated for each test item. The probability of an examinee's answering an item successfully, P(Θ), is plotted as a function of ability (Θ). Items are typically dichotomously scored; receiving a score of 0 or 1. The probability of success is a score ranging from 0 to 1. Ability is scaled as standard scores which theoretically range from $-\infty$ to $+\infty$, but conventionally range from -3 to +3 (see Figure 1). One may interpret the item response curve as: for a given ability Θ, one will have a probability of correctly responding to a given item P(Θ) (Baker, 1983).

The mathematical functions defining ICC's for dichotomously scored items are typically normal ogive curves (Lord, 1980) or logistic functions which yield smooth S-shaped curves. Logistic models will be used in this study because they are most widely used in practical applications of IRT (Costa, 1986) and are easier to interpret. However, it should be noted that logistic functions and ogive functions are mathematically related and provide essentially the same results.

Conceptually, an ICC is similar to a regression function which establishes a mathematical relationship between an examinee's ability level and the probability of success for a given item. Using this analogy, the regression coefficients are known as parameters under IRT.

Figure 1

Item Characteristic Curve



Depending upon the model used, an item characteristic curve may have from one to three parameters:

(a) Item difficulty (b) is the location parameter and determines the position of the ICC along the ability scale. As the curve shifts to the right, the difficulty level of the item increases. Conversely, as the curve shifts to the left, item difficulty decreases. Specifically, "item difficulty represents the point on the ability scale at which an examinee had a 50% probability of answering item *i* correctly" (Hambleton, 1989, p. 154).

(b) Item discrimination (a) is determined by the slope of the ICC at the inflection point (Lord, 1980). Discrimination is the degree to which the response varies with ability level. High values indicate steeper sections of the curve, thus better discriminating power. Low values that generate relatively flat curves have poor discriminating power and are practically useless for ability estimation.

(c) Guessing parameter (c) accounts for the misfit of the ICC at the low ability levels where guessing tends to operate as a significant factor (e.g. academic tests). Known as the guessing parameter, c is employed in the three-parameter logistic models.

In order to generate item characteristic curves, one must estimate item parameters. A model that best fits the circumstances of the test must be chosen (e.g., the two-parameter model). There are three primary models of IRT in the literature:

(a) The one-parameter logistic model (Rasch Model) is characterized by varying item difficulty, holding constant the discrimination of all items. Thus, the curve retains the same shape, but may slide along the ability scale. The one-parameter model assumes no guessing will occur.

$$P_i(\Theta) = \frac{1}{1 + e^{-1(\Theta - b_i)}} \qquad (i = 1,2...n) \qquad (1)$$

where e equals 2.718

b is the difficulty parameter

Θ is the ability parameter

(b) The two-parameter logistic model (Birnbaum, 1968) allows item difficulty and item discrimination to vary. The guessing parameter is held at zero.

$$P_i(\Theta) = \frac{1}{1 + e^{-a_i(\Theta - b_i)}} \qquad (i = 1,2...n) \qquad (2)$$

where a is the discrimination parameter

(c) The three-parameter logistic model is essentially the two-parameter model with the addition of a third parameter "c"; the guessing parameter. This model is typically used in the development of knowledge tests where the estimation of low ability levels is contaminated by guessing error.

$$P_i(\Theta) = c_i + (1-c_i)\frac{1}{1 + e^{-a_i(\Theta - b_i)}} \qquad (i = 1,2...n) \qquad (3)$$

Other models are discussed in the literature (see Hambleton, 1989 for a more complete description of IRT models). Costa (1986) compared five models in analyzing the model of best fit for a motor skill. Those models included the graded item response, partial credit scoring, rating scale, binomial trials, and the Poisson model. Costa used a task analysis of bowling skills with equal discrimination power for all tasks, recommending the use of dichotomous and polychotomous Rasch models for motor skills testing. For the purpose of this study, the one-parameter and the two-parameter models will be compared because the data are dichotomously scored and guessing is not a factor.

Estimating Item Parameters and Ability Scores

Initially item parameters and examinee ability are unknown. Through a process called calibration, item parameters and examinee abilities are estimated and expressed along a metric that provides a frame of reference for interpreting the results, or underlying trait, of the test (Baker, 1985). Test calibration is a two stage process where stage one estimates item parameters and stage two estimates examinee ability. It is a two-stage-closed loop process in which the item parameters and ability estimates are relcalibrated and refined until the set of item functions and ability estimates are stable.

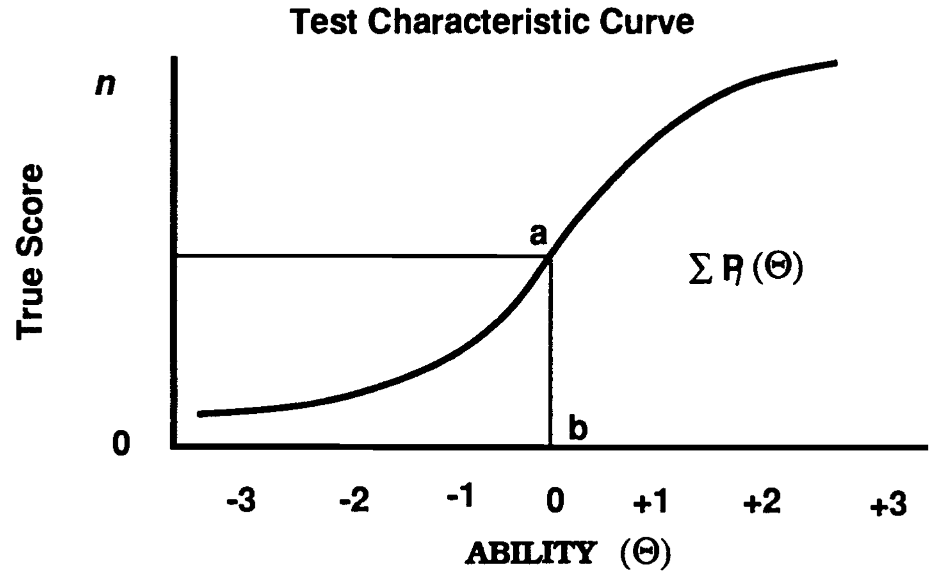
The first stage calculates values for the estimates of item difficulty and discrimination parameters for each item and is referred to as item calibration. At this stage raw test scores are used to estimate ability levels. The second stage employs the values of item parameters calculated in stage one to estimate an examinee's unknown ability parameter. The new ability estimates are then used to recalibrate the item parameters. This iterative process continues until item parameters and ability estimates are stable. The stabilized estimated item parameters are place in the selected ICC equation to compute the $P_i(\Theta)$ at each ability level. Item characteristic curves are plotted with ability estimation on the same metric as the item parameters and represent those that best fit the test data (Baker, 1985).

Item and ability parameters can be estimated using one of several procedures. However, a commonly used procedure is maximum likelihood estimation. The objective of maximum likelihood estimation is to find the value that is most likely for a given set of data. The maximum likelihood estimator (MLE) is an efficient estimator of ability because it is the value with the smallest variance and subsequently less error. The MLE is the combined probability of test items that may be interpreted as the examinee's ability which generates the greatest probability for the observed responses (Hambleton and Swaminathan, 1985). The MLE converges on true values if the sample size and number of items are quite large. Hambleton & Cook (1983) found that 200 or more examinees with test lengths of 20 or more items yielded an acceptable estimated error for the one-parameter model. Higher subject numbers ($n \approx 1000$) have been recommended for more complex models (Hambleton, 1989). The estimated item parameters are invariant across groups (i.e., parameters do not depend on the ability level of examinees responding to the items). The value of item parameters is based on the item, not the group performance. This feature makes IRT more powerful than tests based on CTT, in that tests are not limited to a specific population. However, invariance may not hold true if the item is not used in its proper context; such as measuring more than one latent trait or if the item is too difficult or too easy for a particular population (Baker, 1983).

The Test Characteristic Curve

The test characteristic curve represents a summation of probability of success ($\Sigma P\Theta$) of all n items in a given test battery. It establishes a relationship between true scores and ability level (see Figure 2). The test characteristic curve allows one to transform ability levels into true scores. Thus, it is possible to express an examinee's ability at the true score level or the latent trait level (Θ) (Spray, 1989). The shape of the curve depends on the number of items, the item response model, and values of the item parameters (Baker, 1985).

Figure 2



The test characteristic curve may be interpreted in the same fashion as the item characteristic curve. Test difficulty is determined by the mid-true score value (similar to the P(Θ) = .5 on the item characteristic curve). The slope of the curve is a gross indicator of test score discrimination (Baker, 1985). The test score can be calculated for a single ability level or for a continuum of ability levels. Figure 2 represents a moderately difficult test with a true score of .5 and an ability level that approximates zero. The test curve allows one to directly interpret an ability level as a true score. In addition, the property of invariance holds true for test characteristic curves. The test and item characteristic curves are not dependent on the frequency distribution of a group of examinees.

The Information Function

The contribution of an item does not depend on other items of the test. As a result, item information is a function of the probability of success [P(Θ)] and failure [Q(Θ)] and is defined mathematically for a two-parameter logistic model in equation (4). Information can also be expressed in relation to the the MLE and standard error of estimation [SE(Θ)]. The amount of information (I) given by an item about an estimated ability level is inversely proportional to the variance of the MLE; where I(Θ) = MLE. Information is defined as the reciprocal of the SE(Θ) at a given ability. Larger values of I produce a more precise measurement of ability and smaller confidence intervals around the MLE.

$$I_i(\Theta) = \frac{P_i(\Theta)^2}{P(\Theta)\ Q_i(\Theta)} \qquad (4)$$

$$I_i(\Theta) = SE_i(\Theta)^{-2} \qquad (5)$$

where Q is the probability of an incorrect response.

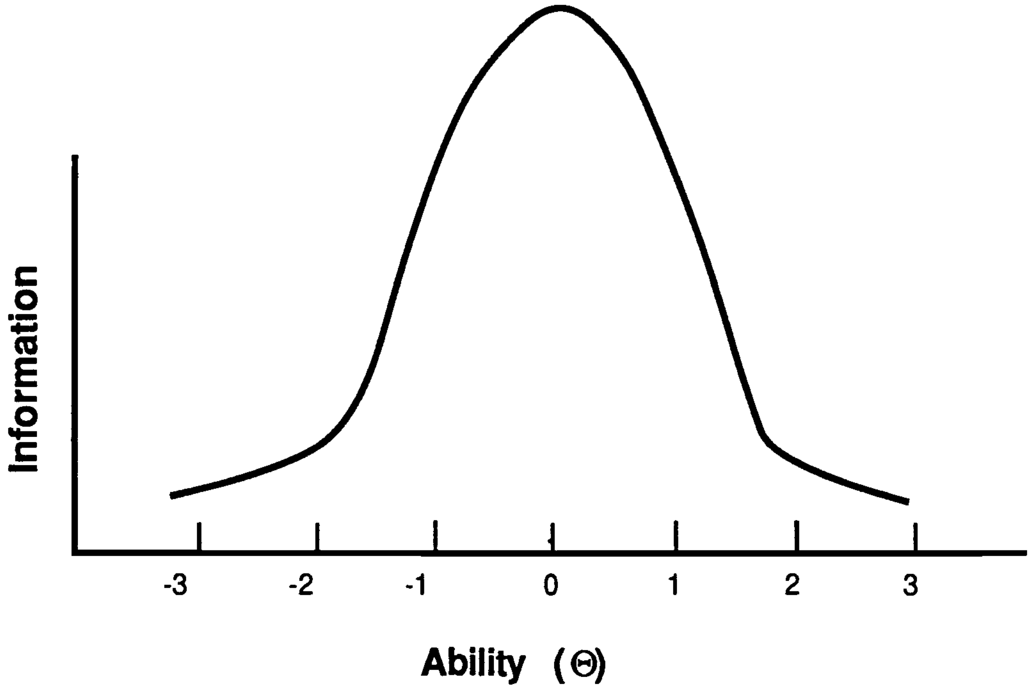
The amount of information possessed by an item is calculated for each ability level and is expected to vary across ability levels. The item information curve plots item information versus ability. Since ability is a continuous variable, the information function is a continuous variable (see Figure 3).

The contribution of individual items is determined independently of other test items (Hambleton, 1989). The item information curve can also determine the ability at which an item provides the best estimation. Item information curves allow the test developer to select items for assessing various ability levels and ranges of ability. Items with relatively low error may be removed from the item bank or test. Since screening and criterion-referenced tests are expected to be most efficient near the mastery level cut-off score, items with information curves that center around the mastery cut-off score are selected for the test. Diagnostic tests, however, contain items that represent a continuum of ability that one is interested in estimating. For these types of tests, item information curves are expected to span the continuum of ability that the test is designed to measure. The item information function allows the test developer to examine the amount of information each item gives at a particular ability level and to determine if the item is appropriate for the test. Item information curves are also employed in adaptive testing where items are chosen to provide the most information near the examinee's ability level.

The amount of information possessed by an item is calculated for each ability level and is expected to vary across ability levels. The item information curve plots item information versus ability. Since ability is a continuous variable, the information function is a continuous variable (see Figure 3).

Figure 3

Information Curve



The test information curve is similar to the item information curve, providing a summation of item information for all items at each ability level [$\Sigma_i I_i(\Theta)$]. The contribution of items to the information of the test is independent of other test items and is additive (Lord, 1980). The test information curve may be particularly useful in validating mastery level cut-off scores for criterion-referenced tests. For example, if the purpose of the test is to screen examinees for special education services, one would: (a) select a low ability level (e.g. $\Theta = -1.5$), (b) administer the test, and (c) identify those who fall below the preestablished cut-off score as qualifying for additional services. One would expect the test information curve to peak at the mastery cut-off score. In contrast, diagnostic test information curves are expected to span a larger area, since they are designed to precisely measure a larger spectrum of ability.

Standard Error of Estimation (SE(Θ))

Information curves provide viable alternatives to classical measures of reliability and standard error (Hambleton & Swaminathan, 1985). Under CTT, error of measurement is the difference between the observed score and the true score (where $X = T + E$). When the true score is fixed, and the error of measurement and observed score have the same standard deviation, that standard deviation is termed standard error of measurement (Lord, 1980). Reliability coefficients and standard error are group dependent and are averaged over all abilities. In contrast, IRT utilizes information functions that are defined independently of a group of examinees. In addition, information may be calculated for any ability level. The SE(Θ) for IRT models serve the same function as the SEM under CTT and can be calculated by summing the $I_i(\Theta)$ values for each item at a given ability level. The SE(Θ) depends on the number of test items, quality of the test items, and the match between item difficulty and examinee ability (Hambleton, 1989). The SE(Θ) is an indicator of the test's precision at ability levels. As with any measurement procedure, the most information can be inferred by the tests with the smallest error. The SE(Θ) allows the test developer to analyze the amount of error for any ability level and for any given set of test items.

Application of IRT to the Psychomotor Domain

Item response theory has had a dramatic effect on measurement in the cognitive and affective domains (Loyd, 1988; Marco, 1977). It has been used by many national test developers of such tests as the Scholastic Aptitude Test (SAT), Graduate Record

Examination (GRE), and has also been employed by state agencies and school districts. Although IRT has been discussed in the literature for the past forty years, it has only gained recent popularity through the availability of computers (Wood, 1987). Application of IRT has grown from simple item and test statistics to encompass test equating, adaptive testing, item banking, pattern analysis, multiple choice test construction, and developing multicomponent models. As research continues, IRT may continue to evolve with the development of additional parameters, models, and new areas of application.

To further examine the potential of IRT models for improvement of measurement and evaluation in the psychomotor domain, Wood (1987) recommended research to include: (a) research papers and discussion of the applicability of IRT to improve tests in the psychomotor domain; (b) examination of IRT literature in the affective and cognitive domains accompanied by analysis of such tests in physical education using IRT methodology; (c) application of current models and/or the development of new IRT models specific to the psychomotor domain; (d) investigation of IRT models and the robustness of the assumptions; and (e) research concerning the practical application of IRT models to motor behavior measurement tools.

Although the foremost application of IRT has been in education and psychology, recent discussions and employment of IRT models to the psychomotor domain offer promise to improve measurement and evaluation in physical education. Item response theory was the focus of a recent Research Quarterly for Exercise and Sport issue (September, 1987). The major issues of applying IRT models to the psychomotor domain focused on (a) the ability of IRT to improve existing measurement and evaluation practices in physical education, and (b) the practicality of meeting the stringent assumptions of IRT with the constraints of testing motor skills. In particular, the assessment of children with handicapping conditions is plagued by the diversity of abilities between and within these populations. For CTT to be effective, tests must be recalibrated for specific groups and abilities. Item response theory is an attractive alternative to CTT because it does not require tests to be recalibrated for different populations unless test items are extremely inappropriate. The invariance characteristic makes IRT very appealing for those who must assess atypical populations because only a few items may be necessary to obtain a valid measure of ability. Empirical research is needed to determine if IRT models can improve traditional methods of assessing motor skills.

Although the application of IRT to psychomotor skill testing may be warranted, the second issue of whether motor skill tests are robust to IRT assumptions, has not been established (Safrit, 1987; Spray, 1987). Motor skill tests are inherently different from knowledge tests in that the testing environment of the latter may be tightly controlled, and therefore changes in performance may be directly attributed to true differences in ability. Motor skill tests, however, are subjected to changes in environmental conditions (Spray, 1987). For example, equipment or facilities may differ, students are typically given more than one trial to perform the skill, and performance is directly observed by an evaluator and/or peers. The fundamental question is whether the conditions under which motor skill tests are administered violate the assumption of local independence. Safrit (1987) contends motor skill testing environments may not violate the invariance assumption because tests may be recalibrated for severe violations of invariance (e.g. different settings). To ensure that the assumption of invariance is not violated, testing conditions should remain fairly consistent. Test developers should clearly state the standard procedures and test administrators should adhere to those procedures. More research is needed to determine if psychomotor tests violate the assumption of invariance.

Empirical application of IRT to the psychomotor domain has been the topic of two dissertations (Charoenruk, 1989; Costa, 1986). Charoenruk (1989) used IRT to equate the Physical Estimation and Attraction Scale for American boys to a Thai version of the same test. Item response theory was useful in this study because each item, as well as the whole test, could be compared. This study provides evidence that IRT may be effective in equating affective test instruments in the psychomotor domain.

Costa (1986) examined the effectiveness of various IRT models in measuring bowling skills. The Rasch model with dichotomous data was compared to polychotomous versions of the Rasch model. Since both models fit the data, Costa concluded that IRT models may be successfully applied to psychomotor skills.

Steffens, Semmes, Werder, and Bruininks (1987) used the Rasch scaling model of IRT to equate the quantitative and qualitative measures of motor development. Item response theory was used to equate scores of the Bruininks-Oseretsky Test of Motor Proficiency (Bruininks, 1979) to The Motor Skills Inventory (Werder & Bruininks, 1988). Item response theory allowed the arrangement of scores along a common scale of difficulty. This enabled the authors to combine qualitative and quantitative motor development data. Ability to determine the contribution of items to a test and invariance across populations makes IRT very powerful and may improve measurement and evaluation in the psychomotor domain. Although the application of IRT to psychomotor skills is relatively new, it may provide a revolutionary approach to measurement in physical education. Since discussions and initial studies show support for the application of IRT models to motor skill tests, it seems reasonable to investigate item and test parameters of current tests.

An application of IRT to the TGMD will provide an opportunity to ascertain the usefulness of applying IRT models to motor skill tests. The present study may provide a basis for future investigations. Test parameters may be recalibrated with atypical populations to develop norms based on ability levels, employ test equating procedures, and pursue the potential of adapted testing. This study follows a prescribed line of research which may assess the potential of IRT to improve measurement and evaluation of motor behavior.

CHAPTER 3

METHODS AND PROCEDURES

The purpose of this chapter is to present the methods, procedures, and statistical analysis proposed for this study. A description of the data set will be presented first, followed by a discussion of the procedures for analyzing the data via IRT.

Subjects

The data were used by Ulrich (1985) to establish the norms for the TGMD. The sample consisted of 913 children (20 mildly mentally handicapped and 897 nonhandicapped) ranging in age from 3 to 10 years. Ulrich used a stratified quota sampling procedure to obtain a sample that represented the 1980 census. Age, gender, race, community size, and geographic region were used as variables to formulate the sample. Statistics describing the sample are available in the TGMD test manual (Ulrich, 1895 p. 24-25). The dichotomously scored data were arranged in rows and columns to represent the performance of the 913 subjects on 43 test items.

Analysis of Data

The TGMD subtests were analyzed separately because a factor analysis (Ulrich, 1985) revealed that the test did not display a simple factorial structure by loading on one dimension. Analyzing the TGMD as one test would have violated the unidimesionality assumption of IRT. Moreover, the division was the most logical and convenient for analysis.

The computer program PC-BILOG (Mislevy & Bock, 1986) was used to estimate IRT parameters and associated statistics. PC-BILOG was designed for a wide range of applications including classical item statistics and IRT parameters. This program was found to give stable and accurate estimates of item parameters and scaled scores (Bock & Aitkin, 1981).

PC-BILOG is organized into three phases. Phase one, the input program, reads formatted data records and calculated classical item statistics. Phase II fits IRT models to each item in the data set. Phase III rescales item parameters and outputs item and test information tables and plots.

Phase I of PC-BILOG: INPUT

The input subprogram calculates selected classical item and test statistics. These statistics include: (a) percent of correct items, (b) the percentage of respondents attempting each item and responding correctly, and (c) item-subscore correlations. Classical item statistics serve as the initial parameter estimates and provide evidence that the data meet IRT assumptions.

Since item parameters of the TGMD are unknown, they were estimated. Estimating item parameters is an iterative process that requires initial estimates. Percent-correct and percentage of correct responses for questions with responses were calculated, printed, and stored for use as the initial estimates for Phase II of the program.

Item correlations are employed to evaluate the assumption of local independence. At a given ability, the pattern of responses should be statistically independent (Hambleton, 1989). Item correlations provide an indication of the items relatedness at a given ability. This does not imply, however, "that test items are uncorrelated over the group of examinees. Positive correlations between pairs of items result whenever there is variation among the examinees on the ability measured by the test items" (Hambleton, 1989, p. 151). If examinees at the same ability level have the same correct responses, the local independence assumption will be violated.

MacDonald (1982) stated that dimensionality should be defined based on the assumption of local independence. To determine the degree of dimensionality at a given ability, score covariance between items in the test should be zero. "The dimensionality of a set of test items is defined as the number of traits needed to satisfy the principle of local independence" (Hambleton, 1989, P. 151). Nonlinear factor analysis should be used since the covariance of the items is usually nonlinear.

Phase II of PC-BILOG: CALIBRATE

Phase II fits a logistic item-response function to each item of the test. The one- and two-parameter logistic IRT models were compared to determine which model best fit the data. Estimation of the one-parameter model is less complex since item discrimination is held constant. In contrast, both item discrimination and item difficulty must be estimated for the two parameter model. Initial item-parameter values were estimated from proportion correct (item difficulty) and biserial correlations (item discrimination) derived in Phase I.

The marginal maximum likelihood estimator (MMLE) was used for both the one-parameter and the two parameter model to estimate the item parameters. The maximum likelihood function was not obtainable since the true item parameters were unknown. Hambleton described the MMLE as an attractive alternative estimation procedure because it "removes ability estimates from the estimation of item parameters by integrating them out of the likelihood function. The resulting likelihood function is used to find item-parameter values that maximize the marginal likelihood function" (Hambleton, 1989, p. 170). The MMLE is a useful procedure because item parameters may be calculated without estimating the ability parameter (Lord, 1986).

In Phase II item difficulty, item discrimination (slope), and plots of expected item and subtest curves were derived. Each examinee was assigned to one of 20 intervals that contained the maximum likelihood estimate of their score. In addition to item parameters, Phase II provides chi-square indices of fit for each item upon completion of the final item estimation cycle. The expected proportion for each interval was calculated with the average expected response probabilities of subjects in the respective interval. Extreme intervals were combined so that expected frequencies exceed four. BILOG plots observed and expected item-response curves for each item.

Phase III of PC-BILOG: SCORE

The SCORE subprogram calculates ability from the item statistics from Phase I and item parameters calculated in Phase II. The investigator was given the option to select from the maximum likelihood, Bayes model, maximum a posteriori, or the Bayes expected a posteriori for calculating ability. Item parameters of very easy or very difficult items may be recalculated based on the model chosen. These models account for the error inherent in estimating items with extreme difficulty parameters.

Initially, ability estimates were scaled to the same metric as the item parameters. Phase III allowed the investigator to rescale ability estimates from Phase II to display items on a desired ability scale for tests or subtests. Standard errors and item parameters were rescaled to match the rescaled ability estimates.

The following plots and indices were printed to examine the psychometric properties of the TGMD:

(a) tables of traditional item statistics and IRT item parameters;

(b) tables of item information, including point and maximum information value, and corresponding standard errors;

(c) tables of ability estimates and standard errors;

(d) plots of character function and information function for each item;

(e) subtest information and standard error curves.

Assessment of Model Fit

The one-parameter and the two-parameter models were compared to determine which model best fit the data set. It is important that the model fit the data because poorly fitting models do not yield accurate ability and true score estimates. Moreover, these models cannot claim the invariance property of IRT.

Chi-square tests were used to assess model fit. A scaled score continuum was divided into successive intervals for displaying response proportions (Θ). Examinees were grouped into categories based on their ability estimates; the maximum number of categories was 20 in the BILOG program. Observed and expected response curves were fit to the data. The chi-square statistic, relative to the product of the binomial alternative was calculated (Phase II). Significant chi-square values at the $p < .05$ indicated a poor fit of the item to the IRT model.

Assessment of the Psychometric Properties of the TGMD

The following item response characteristics were compared to the appropriate CTT statistics:

(a) the difficulty parameter for each item, separately for each subtest;

(b) the discrimination parameter for each item, separately for each subtest; and

(c) the information function for each item, separately for each subtest.

Standard Error of Estimation

The standard error of measurement (SEM) in CTT and the standard error of estimation in IRT (SE(Θ)) are fundamentally similar. The SEM and SE(Θ) provide an index of the test's reliability and a confidence interval in which a student's true score (or ability) is expected to lie. These values are reported on the same scale as the reported examinee's test or ability score. The SEM is based on the error of measurement associated with the test and is averaged across subjects and ability levels. Conversely, the SE(Θ) is associated with the estimation of ability and is not averaged across subjects or ability levels. The SE(Θ) is the reciprocal of the square root of the information function.

$$SE_i(\Theta) = I_i(\Theta)^{-1/2} \qquad (i= 1,2,\ldots n) \qquad (5)$$

The $SE_i(\Theta)$ was examined to determine the relative precision of the TGMD in estimating ability from test items and subtests. Ulrich found the subtests' SEM to be acceptable for measuring gross motor development for children ranging from 3 to 10 years of age. It was expected that all items would have the same SE(Θ) and that the $SE_i(\Theta)$ would vary with ability level. The SE(Θ) for both subtests were compared to determine the precision of these measures at various ability levels. Information functions and SE(Θ)'s provided a general indicator of the ability range best measured by the TGMD.

Mastery Classification

Mastery Classification Agreement

The purpose of mastery classification is to determine if an examinee has reached a predetermined achievement or performance level. Minimum level of mastery is chosen by considering the consequences of misclassifying a master or a nonmaster. Once the classification criterion has been chosen, a test is analyzed for its decision accuracy. A contingency table is drawn that compared the true state of mastery with the classification decision based on the test results. The true state of mastery may be determined by a number of methods (see Safrit, 1989 for a description of these methods). Item response theory true score estimates were used in the present study to represent the true mastery state. This theory provides an estimation of examinees' latent trait and true score. It is not assumed that IRT is the absolute measure of true ability, but it provides a more precise estimation of true ability. It should be noted, however, that IRT estimation of the "true" mastery state, like other estimation methods, suffers from a degree of subjectivity in choosing the most valid cut-off score.

Ulrich analyzed the reliability of 70% and 85% mastery level cut-off scores for classifying nonhandicapped and handicapped children, but did not examine the validity of these mastery level cut-off scores. These levels were chosen based on Popham's (1978) recommendation and were found to have acceptable test-retest reliability. The present study examined the mastery cut-off score at true scores that represented 70% and 85% mastery.

The mastery classification scores were computed for CTT raw scores and IRT ability scores. Raw scores were calculated by multiplying 70% and 85% by the number of items in each subtest. The IRT cut-off scores were found by the following procedure:

1. The true score that represents 70% and 85% mastery for each subtest was determined by using the probability of success statistics.
2. The mastery level cut-off true score was transformed into an ability score (Θ_O) by using the test characteristic curve equation and the item characteristic curve equation for the two-parameter model (Baker, 1985). The following equation was to estimate Θ_O:

$$(\Theta_O) = \Sigma P_i(\Theta);\text{ where } P_i(\Theta) = 1 + [1 + e^{-1(\Theta - b_i)}] \qquad (7)$$

where e = 2.718; *b* is the difficulty parameter; and (Θ) is the ability parameter.

The validity of the mastery level cut-off scores was analyzed in two parts. The first part examined the agreement of mastery classification between CTT and IRT analyses for each subtest. Contingency tables were drawn to assess the agreement of

Figure 4

Contingency table comparing the mastery classification decisions based on CTT and IRT

		Item Response Theory	
		Master	Nonmaster
Classical Test Theory	Master	Correct Classification (cell a)	Incorrect Classification (cell b)
	Nonmaster	Incorrect Classification (cell c)	Correct Classification (cell d)

According to Safrit (1981), the contingency coefficient (C) is calculated by the following steps:

1. C is the proportion of agreement between the data sets which is calculated by summing the diagonal master/master and the nonmaster/ nonmaster cells.

2. Kappa is reflects the agreement of mastery classification that was due to chance: $k = (C - Pc)/ 1- Pc$ where $Pc = \Sigma(P_i)(P_j)$. and P_i is the summed percent across rows or columns, for example (see Figure 4):

   Pc = (cell a + cell b)(cell a + cell c) + (cell c + cell d)(cell b + cell d)

Precision of Mastery Classification

The second part of evaluating the mastery level cut-off scores entailed an examination of information curves to determine the precision of the subtests to estimate ability at the level that represented the cut-off scores. The primary goal was to evaluate subtest information functions at ability levels that represented 70% and 85%

estimate ability at the level that represented the cut-off scores. The primary goal was to evaluate subtest information functions at ability levels that represented 70% and 85% mastery and report the amount of information supplied at that ability level. The maximum point of information was not expected to center at the 70% and 85% mastery levels. Thus, the point at which the maximum point of information centered and the mastery level it represented is reported in chapter 4 of this study.

CHAPTER 4

RESULTS AND DISCUSSION

Introduction

Chapter IV consists of three sections. In section 1, a summary of classical test theory (CTT) analysis with traditional item statistics is presented. The item response theory (IRT) analysis of the TGMD is given in section 2. Section 3 provides a brief discussion of the advantages and disadvantages of using IRT for analyzing the Test of Gross Motor Development (TGMD).

The original data that were used to establish the TGMD norms (Ulrich, 1985) were modified for this study to accommodate the restrictions of the IRT analysis. Thirty-two subjects were deleted from Ulrich's orginal data because they performed at 100% mastery; no subjects in the data set exhibited 0% mastery. Item response theory cannot provide accurate estimates of ability at mastery levels of 0% and 100% because this suggests that the test is either too easy or difficult for the subjects (Hambleton, 1989). Consequently, the interpretation of the IRT analysis is only valid for those who exhibited a raw score of 44 to 1 on the TGMD. Deleting those exhibiting perfect scores may have slightly biased (i.e. underestimated) the statistics reflecting agreement between CTT and IRT in classifying masters. However, since only 32 subjects out of 913 (3.5%) were deleted, the change in the results is negligible.

The TGMD subtests were analyzed separately because a factor analysis (Ulrich, 1985) revealed that the test did not display a simple factorial structure by loading on one dimension. Analyzing the TGMD as one test would have violated the unidimensionality assumption of IRT. Moreover, the division was the most logical and convenient for analysis.

Traditional Item Analysis

The PC-Bilog computer program (Bock & Mislevy, 1986) was used to generate item and subtest parameters. Phase I of the PC-Bilog program calculated traditional item difficulty and item discrimination statistics. These statistics are in reference to the standardized sample of the TGMD and are not invariant across populations. Table 3 presents the traditional CTT item statistics for the TGMD subtests.

Table 3
Traditional Item Analysis Statistics for the Locomotor and Object Control Subtests

	Locomotor			Object Control	
Item	Difficulty	r_b	Item	Difficulty	r_b
1	.995	.400	27	.856	.498
2	.825	.699	28	.690	.588
3	.878	.488	29	.427	.864
4	.730	.598	30	.424	.754
5	.795	.626	31	.501	.874
6	.954	.499	32	.479	.732
7	.424	.483	33	.569	.852
8	.794	.483	34	.793	.443
9	.769	.903	35	.773	.753
10	.533	.868	36	.506	.858
11	.470	.733	37	.501	.967
12	.785	.755	38	.897	.650
13	.575	.819	39	.409	.891
14	.553	.789	40	.426	.869
15	.313	.721	41	.237	.667
16	.699	.737	42	.665	.833
17	.307	.629	43	.475	.914
18	.957	.277	44	.619	.766
19	.655	.757	45	.353	.811
20	.679	.899			
21	.607	.891			
22	.520	.877			
23	.669	.524			
24	.811	.571			
25	.833	.483			
26	.761	.619			

Item difficulty of the locomotor subtest ranged from .307 to .995 (median = .715, SD = .186). Items 1, 2, 3, 6, 18, 24, and 25 were very easy as indicated by difficulty values higher than .800. Items 1, 2, and 3 represented the run test which appeared to have a very low degree of difficulty. The horizontal jump contained one easy item (18). Items 24 and 25 represented easy items of the skip. The object control subtest difficulty values ranged from .237 to .897 (median = .501, SD = .180). Item 27 of the two-handed strike and item 38 of the kick displayed a difficulty value higher than .800. In contrast, item 41 of the kick was the only item on the TGMD that exhibited difficulty value less than .300. Preliminary examination of traditional item difficulty suggests that the object control subtest was more difficult than the locomotor subtest.

Item discrimination, represented by the biserial correlation statistic (r_b), is defined as the correlation between responses to an item and the subtest total score. High values indicate that subjects who master item *i* tend to score higher on the subtest, while subjects who fail to master the item tend to achieve a low subtest score. Although interpretation of item discrimination values varies with test purposes, a general indicator of item discrimination for most test purposes is as follows: ≥ .4 represents very good or very high discrimination, .3 - .39 provides high discrimination, .2 - .29 indicates marginal discrimination, and values < .2 represent unacceptable discriminating power (Ebel, 1965).

Discrimination values for the locomotor subtest ranged from .277 to .903 (median = .71, SD = .168). The locomotor subtest in general displayed very good discrimination. One item (item 18) on the locomotor subtest had an item discrimination value lower than .3. Conversely, all items on the object control subtest displayed high or very high discrimination values that ranged from .443 to .967 (median = .811, SD = .142). Traditional item statistics suggest that the object control subtest displayed better discriminatory power than the locomotor subtest.

IRT Analysis

Model Fit

To assess the fit of IRT models, the PC-Bilog program calculates a goodness-of-fit index. A chi-square statistic was computed for each item on the TGMD to compare goodness-of-fit of the one-parameter and two-parameter models. A significant chi-square value indicates a poor fit of the model for a given item. The advantages of IRT can be claimed only if the data fit the item response model. A poorly fit model will not yield invariant parameters and ability estimates (Hambleton, 1989). Misfit items may be due to (a) employing the wrong item response model or (b) the values of the proportion of correct responses vary greatly in the sample, thus making the fit of any model virtually impossible. Most tests will yield a few items that do not fit the model due to the second reason (Baker, 1985). However, if many items do not fit the model, the test developer can be confident that the model does not fit the data.

As expected, neither the two-parameter nor the one-parameter model fit the TGMD data perfectly. The two-parameter model yielded a more accurate fit than the one parameter model, as indicated by the number of items poorly fitting the models. The one-parameter model poorly fit 17 locomotor (items 1, 3, 4, 5, 8, 9, 10, 11, 12, 13, 18, 20, 21, 22, 23, 24, and 25) and 11 object control items (27, 28, 29, 31, 34, 35, 37, 39, 40, 41, and 43). In contrast, the two-parameter model revealed significant chi-square statistics for only three locomotor (items 1, 8, and 20) and five object control items (items 28, 29, 32, 35, and 38). Items that yielded significant chi-square values under the two-parameter model were also significant with the one-parameter analysis. Clearly the two parameter model fit the TGMD data more precisely than the one-parameter model, indicating that the item discrimination index was not equivalent for all items on the TGMD.

Werder and Bruininks (1987) used the one-parameter model to equate the scores of the Bruininks-Oseretsky Test of Motor Proficiency (Bruininks, 1978) and The

Motor Skills Inventory (Werder & Bruininks, 1988). The Bruininks-Oseretsky Test of Motor Proficiency provides a quantitative index of fine and gross motor proficiency while the Motor Skills Inventory (1988) gives a qualitative measure of fundamental motor skills and movement patterns. An IRT analysis enabled the authors to combine qualitative and quantitative motor development data to express the scores along the same scale of difficulty. Werder and Bruininks chose to employ the one-parameter model of IRT. The one-parameter model differs from the two-parameter model in that item discrimination of the one-parameter model is held constant, while item discrimination of the two-parameter is free to vary. The authors did not discuss the reasons for choosing the one-parameter model nor did they discuss the fit of the one-parameter model. Since the Bruininks-Oseretsky Test of Motor Proficiency, the Motor Skills Inventory, and the TGMD test psychomotor skills, one would expect similar results regarding model fit. It is very unlikely that all items on motor skills tests have equal discrimination indices. Moreover, the analysis of model fit should investigate more than one model to determine the model of best fit (Hambleton, 1989; Baker, 1985). Future studies which apply IRT to psychomotor skills should investigate more than one model (i.e. one-parameter, two-parameter, or multidimensional models) to determine the model of best fit.

IRT Comparison By Subtest

For the given data set, interpretation of IRT difficulty and discrimination parameters parallels interpretation of CTT item statistics. The correlation between traditional and IRT item difficulty was expected to yield a moderate-high negative relationship because traditional difficulty values are inversely proportional to the degree of difficulty and IRT item difficulty increases with the degree of difficulty. Conversely, a moderate-high positive correlation is expected between CTT item discrimination and IRT discrimination because larger values indicate better discrimination under both test theories.

The Pearson Product Moment correlation was used to assess the linear relationship between the item parameter values computed by CTT and IRT. As expected, traditional item difficulty of the locomotor subtest revealed a high negative correlation (r = -.869) with the IRT values. Object control subtest difficulty displayed a very high negative correlation (r = -.981). The locomotor subtest discrimination indices correlated highly (r = .910), while the object control subtests revealed a slightly higher linear relationship (r = .938). As expected, IRT and traditional analysis provided parallel interpretations regarding item analysis.

It is encouraging that the interpretations of the CTT statistics and IRT parameters were closely related because it provides evidence for the validity of the IRT analysis. If the IRT analysis did not provide similar interpretations, one would question the suitability of the TGMD data for an item response analysis. However, since CTT and IRT revealed parallel interpretations of item statistics and item parameters, to what advantage is the employment of IRT over CTT?

Classical test theory currently serves as the foundation for most tests in the psychomotor domain, is better known by researchers and professionals in the field, and does not require very large sample sizes nor stringent assumptions regarding test scores and underlying ability. However, interpretation of CTT test scores is limited to a defined population and intended uses. Practitioners are restricted to assess only those who meet the characteristics of the defined population. This creates problems for those who must assess a wide variety of abilities because most tests are designed to measure higher ability levels then those found in special education programs. Moreover, CTT cannot determine the error or precision of each test item at various ability levels because error is assumed to remain constant across all items and ability levels. Thus, it is nearly impossible to make precise estimates of the error in estimating an examinee's ability. Classical test theory is further limited in equating test scores or determining if a test is bias in measuring different populations. For example, a test developer may wish to determine if a test accurately estimates ability for a nonhandicapped and handicapped population. CTT could not equate the two tests because traditional item statistics and interpretations are limited to one population, thus comparisons among populations and tests are not valid. Item response theory produces invariant item parameters which enable researchers to make valid comparisons among tests to examine test bias and equate tests.

Item response theory compliments and enhances CTT. Although stronger assumptions must be met, larger sample sizes tested, and somewhat nebulous scores interpreted, IRT displays distinct advantages over CTT to assess a given underlying ability or trait. Item response theory can be used to develop tests that meet the needs of practitioners and satisfy the criteria of well designed tests. Item response theory provides parameters that are not population bound, freeing the practitioner to assess a wider variety of populations and abilities. The precision of each item to measure the ability continuum may be determined. Thus, tests may be constructed to measure low abilities and effectively measure change in a wide range of the ability continuum.

Furthermore, knowledge of item parameters is useful for employing other forms of testing and optimally improving tests and test administration in the motor domain. One avenue of potential application is adaptive testing. Adaptive testing may improve the efficiency of motor skill testing because only those items that approximate an examinee's ability are administered. Thus, practitioners save administration time while maintaining the motivation of examinees by presenting challenging items that are of appropriate difficulty. Maintaining the motivation of examinees is difficult, especially when a high rate of failure is incurred. The test designer can use item parameters to arrange items on a continuum of difficulty. Easy items may be presented first, providing success and consequently motivating the examinee. As more difficult items are presented, the examinee is challenged and interest is maintained. Upon successive failure, the test may be terminated and an ability may be estimated. There are many more potential applications of IRT to measurement in the psychomotor domain. The results indicating parallel interpretation between CTT and IRT test items are encouraging and lend support for further investigation of IRT in the assessment of motor skills.

IRT Subtest Analysis

The locomotor subtest IRT difficulty parameters ranged from -6.23 to 0.743 (median = -0.944, SD = 1.733). Three items of the locomotor subtest (items 1, 6, and 18) displayed very low difficulty in comparison to other items on the subtest. Object control difficulty values ranged from -2.097 to 1.133 (median = .053, SD = .895). The object control subtest displayed a higher average degree of difficulty than the locomotor subtest.

Baker (1985) provided the following general guidelines for the interpretation of IRT item discrimination indices: values > 1.7 as having very high discriminatory power, 1.35 - 1.69 denote high discrimination, .65 - 1.34 have moderate power, .35 - .64 reveal low discrimination, and indices < .01 have no discrimination. Locomotor discrimination values ranged from .578 to 2.963 (median = 1.54, SD = .732). One item (item 18) displayed a discrimination index lower than .65. This item was the only item to have poor discriminatory power under traditional item analysis. Discrimination values for the object control subtest ranged from .85 to 3.548 (median = 2.17, SD = .663). In general, the object control subtest exhibited higher discriminatory power than the locomotor subtest by traditional and IRT analyses.

Subtest Information Function

The information function is an indicator of the precision with which the subtest item estimates ability, where $I = SE(\Theta)^{-2}$. As item information increases, the standard error of estimation decreases. Ideally, an information curve would provide a large magnitude of information across a wide range of abilities, displaying a high and wide curve. Since items usually assess only a small portion of the ability continuum, typical information curves are peaked and centered near or slightly above the difficulty parameter. Information curves are difficult to analyze because there are no general guidelines currently available for interpretation. Furthermore, information values are dependent on the model selected and number of items comprising the test (Hambleton, 1989).

Information curves are described by the maximum amount or value of the curve and the point at which the maximum value lies on the ability scale. The point of maximum information is the ability level (ϴ) at which the subtest/or item estimates the underlying trait with the least amount of error. The value of maximum information (I) is the value at the highest point of the information curve. On a scale from 1 to 19, the two subtests yielded high maximum information values (see Appendix C). The value of maximum information for the locomotor subtest indicated that the subtest provided a high degree of precision (I = 15.50) at a low ability level, where the point of maximum information was at ϴ = -1.857. The subtest information curves are a sum of the item information curves. Item information values of the locomotor subtest ranged from .084 to 2.323 (median = .593, SD = .656). The object control subtest provided more information I = 18.24 at a slightly higher ability level (ϴ = -1.643), where item information values ranged from .181 to 3.147 (median = 1.178, SD = .506). In general, the object control subtest was a more precise measure of ability. Maximum information values for both tests provided evidence that the TGMD best measures very low ability with adequate precision.

Standard Error

The standard error of measurement (SEM) in CTT and the standard error of estimation in IRT (SE(ϴ)) are conceptually similar. The SEM and SE(ϴ) provide an index of the test's precision that can serve as a confidence band which spans a student's true score. These values are reported on the same scale as the reported examinee's raw test score (SEM) or ability score (SE(ϴ)). The SEM is based on the error of measurement associated with the test and is averaged across subjects and ability levels. Conversely, the SE(Θ) is associated with the estimation of ability for individual subjects and is not averaged across subjects or ability levels. The SE(Θ) depends on the number of test items, quality of test items, and the match between item difficulty and examinee ability (Hambleton, 1989).

The SE(Θ) of the TGMD could not be computed because the test was analyzed by subtests. Combining the subtest values would have violated the IRT assumption of unidimensionality. The SEM for the locomotor subtest was .60 and the object control subtest revealed an SEM of .52. The SE(Θ) at I of the locomotor subtest was .25. The SE(Θ) at I of the object control subtest was calculated as .30. The SE(Θ) changes with ability level as exhibited in Appendix C.

Item Analysis

For the purposes of discussion, the items were grouped into the 12 motor skills that comprise the TGMD. Item discrimination, item difficulty, and the maximum value of information (I) will be interpreted for each item. The point of maximum information will not be presented because it usually approximates the value of the difficulty parameter (*a*). Item parameters are provided in Table 3, while item characteristic and information curves are grouped as skills and presented in Appendix B.

Table 3

IRT Item Analysis

Skill	Item	*b* parameter	*a* parameter	*I*
Run	1	-6.230	0.964	0.2323
	2	1.548	1.469	0.5398
	3	-2.659	0.861	0.1864
	4	-1.198	1.083	0.2930
Gallop	5	-1.556	1.152	0.3321
	6	-3.981	0.761	0.1927
	7	0.272	1.984	0.9842
	8	-2.058	0.761	0.1447
Hop	9	-0.989	2.564	1.6438
	10	-0.082	2.919	2.1296
	11	0.112	1.889	0.8918
	12	-1.291	1.498	0.5608
Leap	13	-0.246	2.295	1.3163
	14	-0.188	1.921	0.9224
	15	0.633	2.415	1.4587
Horizontal Jump	16	-0.813	1.645	0.6764
	17	0.743	1.654	0.6842
	18	-5.710	0.578	0.0836
	19	-0.629	1.580	0.6244
Skip	20	-0.613	2.522	1.5900
	21	-0.341	2.677	1.7917
	22	-0.040	2.963	2.1951
Slide	23	-0.898	0.998	0.2490
	24	-1.846	0.982	0.2409
	25	-2.348	0.792	0.1594
	26	-1.401	1.066	0.2841
Locomotor Subtest				15.500
Two-hand Strike	27	-2.097	1.035	0.2677
	28	-0.882	1.153	0.3322
	29	0.305	2.424	1.4689
	30	0.335	1.746	0.7625
Stationary Bounce	31	0.053	2.300	1.3230
	32	0.128	1.583	0.6263
	33	-0.191	2.261	1.2779
Catch	34	-1.828	0.850	0.1808
	35	-1.061	1.896	0.8983
	36	0.036	2.249	1.2645
	37	0.068	3.548	3.1474
Kick	38	-1.943	1.637	0.6700
	39	0.363	2.836	2.0110
	40	0.312	2.476	1.5331
	41	1.133	1.642	0.6744
Overhand Throw	42	-0.555	2.170	1.1777
	43	0.147	2.754	1.8967
	44	-0.416	1.739	0.7539
	45	0.571	2.268	1.2858
Object Control Subtest				18.240

Locomotor Subtest

The run included four items:

Item 1: "brief period where both feet are off the ground." This item is not very useful for measuring motor development. Item 1 was confounded by a significant chi-square value which indicated a misfit of the model to this item. Thus, item parameters and information values are not accurate

Item 2: "arms in opposition to legs, elbows bent" was an easy item (*b* = -1.548) with good discrimination (*a* = 1.469) and a maximum information value of .540.

Item 3: "foot placement on or near a line (not flat footed)" was also an easy item (*b* = -2.695) with moderate discrimination. It displayed a low magnitude of information (I = .186).

Item 4: "nonsupport leg bent approximately 90 degrees (close to buttocks)." This item had a low degree of difficulty (*b* = -1.198), moderate discriminatory power (*a* = 1.083), and supplied a low degree of information (I = .293).

The run was an easy skill. Most of the items had moderate discrimination and item 2 exhibited high discrimination. Items 1 and 3 did not adequately fit the two-parameter model, as indicated by significant chi-square statistics. Because of these poor fitting items and very low degree of difficulty, the run should be revised.

The gallop was represented by items 5 through 8:

Item 5: "a step forward with the lead foot followed by a step with the trailing foot to a position adjacent to or behind the lead foot" was an easy item (*b* = -1.556) with good discrimination (*a* = 1.152) and provided a moderate-low amount of information (I = .332).

Item 6: "brief period where both feet are off the ground" was an extremely easy item (b = -3.981) with moderate discriminatory power (a = .878) and low information (I = .193).

Item 7: "arms bent at waist level" exhibited classic item parameters, with slightly above average difficulty (b = .272), very good discriminating power (a = 1.984), and moderate information (I = .984).

Item 8: "able to lead with the right and left foot" did not fit the two-parameter model as indicated by a significant chi-square statistic. Thus, item parameters and information values are inaccurate and item 8 should be revised.

The gallop measured a wide range of ability. Items 5, 6, and 8 were very easy, while item 7 exhibited good psychometric properties for measuring average ability. Item 5 would be useful in measuring low abilities because of its low b parameter and high discrimination. Item 6 is of marginal use since it exhibited very low difficulty and information. Item 8 showed little utility since it did not fit the model.

The hop incorporated items 9 through 12:

Item 9: "foot of nonsupport leg is bent and carried in back of the body" was of moderate-high difficulty (b = -0.989) and exhibited very good discriminatory power (a = 2.564) and information (I = 1.644).

Item 10: "nonsupport leg swings in pendular fashion to produce force" best measured moderate ability (b = -0.082). It discriminated very well (a = 2.919) and provided a large amount of information (I = 2.13).

Item 11: "arms bent at elbows and swing forward on take off." This item had very high discriminatory power (a = 1.889) and provided the most information (I = .892) at medium difficulty (b = 0.112).

Item 12: "able to hop on right and left foot" best measured low ability (*b*= -1.291). It exhibited very high discrimination (*a* = 1.498) and a moderate amount of information (I = .561).

The easiest item of the hop was item 12 followed by items 9, 10, and 11. The hop contained items that, in general, exhibited good psychometric properties which discriminated very well among those of average ability. The items provided acceptable information, and fit the two-parameter model.

The leap contained three items; items 13 through 15:

Item 13: "take off on one foot and land on the opposite" exhibited good psychometric properties. It measured moderate ability (*b* = -0.246) with very high discriminatory power (*a* = 2.295), and exhibited high information (I = 1.316).

Item 14: "a period where both feet are off the ground (longer than running)." This item displayed moderate difficulty (*b* = -0.188), discriminated very well (*a* = 1.921), and supplied a moderate-high amount of information (I = .922).

Item 15: "forward reach with arm opposite of lead foot" was one of the more difficult items (*b* = .633). It had very good discriminatory power (*a* = 2.415) and provided a large magnitude of information (I = 1.459).

The items of the leap discriminated very well at average or slightly above average (item 15) ability. Item 13 was the easiest followed by items 14 and 15. The leap exhibited good psychometric properties and all items fit the two-parameter model.

The horizontal jump encompassed items 16 to 19:

Item 16: "preparatory movement includes flexion of both knees with arms extended behind the body" best measured subjects with moderate-low ability (*b* = -0.813) with very high discriminatory power

(a = 1.645) and moderate information near the lower ability levels (I = .676).

Item 17: "arms extended forcefully forward and upward, reaching full extension above the head." This item was one of the more difficult items (b = 0.743) and exhibited very good discriminating power (a = 1.654). Item 17 provided moderate-high amount of information (I = .684). Item 17 is a good item for estimating higher ability levels.

Item 18: "take off and land on both feet simultaneously." Item 18 exhibited psychometric properties of low difficulty (b = -5.710), marginal discrimination (a = .578), and poor information (I = .084). This item and item 1 of the run were both much easier than the other items on the test. Item 18 should be revised.

Item 19: "arms are brought downward during landing." This item measured a moderate-low ability level (b = -0.629) with very high discriminating power (a = 1.580) and supplied a moderate amount of information (I = .624).

The horizontal jump contained items that exhibited good psychometric properties (items 16, 17, and 19) and one item that should be revised (item 18). Items 16, 17, and 19 had excellent discrimination values and assessed a wide range of the ability continuum, by measuring moderate-low ability (Θ = -.813) to moderate-high ability (Θ = .734). The easiest item was 16, followed by 19 and 17. Item 18 evidenced low utility by estimating very low ability with unacceptable precision.

The skip was represented by three items; items 20 through 22:

Item 20: "a rhythmical repetition of the step-hop on alternate feet" is of little utility because it did not fit the two-parameter model.

Item 21: "foot of nonsupport leg is carried near surface during hop". This item best measured slightly below average ability (*b* = -.341). Its discriminatory power was very good (*a* = 2.677) and supplied a large amount of information (I = 1.792).

Item 22: "arms alternately moving in opposition to legs at about waist level" displayed moderate difficulty (*b* = -.040) and had extremely high discrimination (*a* = 2.963) and information (I = 2.195).

The skip contained items that exhibited good psychometric properties by discriminating well among those of moderate-low to average ability except for item 20 which did not fit the two-parameter model and should be revised.

The slide is comprised of items 23 though 26:

Item 23: "body turned sideways to desired direction of travel" best functioned at low ability levels (*b* = -.898). This item had moderate discriminatory power (*a* = .998), but provided a low amount of information (I = .249).

Item 24: "a step sideways followed by a slide of the trailing foot to a point next to the lead foot" measured low ability (*b* = -1.846). It displayed moderate discrimination (*a* = 0.982) and a low information curve over a wide range of ability levels (I = .241).

Item 25: "a short period where both feet are off the floor" best measured very low ability (*b* = -2.348) with moderate discriminatory power (*a* = 0.782). Its maximum information value was low (I = .159), and thus did not precisely estimate ability.

Item 26: "able to slide to the right and to the left" was an easy item (*b* = -1.401) with moderate discrimination (*a* = 1.066) and suppled a moderate-low amount of information (I = .284).

The slide was an easy skill, where item 25 was the easiest followed by item 24, 26, and 23. However, interpretation of ability is limited by the low amount of information exhibited by all of the items (I-values ranged from .159 to .284).

Object Control Subtest

The two-hand strike was the first skill of the object control subtest. It included items 27 through 30:

Item 27: "dominant hand grips bat above nondominant hand" was a very easy item (*b* = -2.097) with moderate discrimination (*a* = 1.035) and moderate-low precision (I = .267).

Item 28: "nondominant side of body faces the tosser (feet parallel)." This item did not fit the two-parameter model and thus did not provide accurate item parameters and information values.

Item 29: "hip and spine rotation." This item did not fit the two-parameter model and should be revised.

Item 30: "weight is transferred by stepping with front foot". This item was a fairly difficult item (*b* = .335) that had very high discriminatory power and provided good information (I = .763).

The two-handed strike items should be revised since two out of three items did not fit the two-parameter model, and the first item displayed only marginal precision.

The stationary bounce contained three items; items 31 through 33:

Item 31: "contacts ball with one hand at about hip height" exhibited moderate difficulty (*b* = .053), high discrimination (*a* = 2.30), and displayed a steep information curve (I = 1.323).

Item 32: "pushes ball with fingers (not a slap)" did not fit the two-parameter model, thus interpretation of item parameters and information values is inaccurate.

Item 33: "ball contacts floor in front of or at the side of the hand being used". This item measured medium difficulty (*b* = -0.191) with very good discrimination (*a* = 2.261) and supplied a large amount of information (I = 1.278).

The items of the stationary bounce had very high discriminatory power to estimate moderate ability. Although the items seemed to display good psychometric properties, item 32 did not fit the two-parameter model and should be revised.

The catch was represented by items 34 through 37:

Item 34: "preparatory phase where elbows are flexed and hands are in front of body." This was an easy item (*b* = -1.828) with moderate discrimination (*a* = .850) and low information (I = .181).

Item 35: "arms extended in preparation for ball contact" confounded by a lack of model fit. Item 35 should be revised.

Item 36: "ball is caught and controlled by hands only" displayed medium difficulty (*b* = .036) and very high discrimination (*a* = 2.249). Its information curve was high, indicating good precision (I = 1.265).

Item 37: "elbows bent to absorb force" was slightly more difficult than item 36 (*b* = .068). Its discriminatory power was extremely high (*a* = 3.548) and the item provided a large amount of information (I = 3.147).

The catch contained items that were relatively easy. Item 34 was the easiest, followed by items 35, 37, and 36. All items had moderate (item 34) or high discriminatory power. Although the items of the catch collectively seemed to assess low to average ability fairly well, items 34 and 35 poorly fit the two-parameter model. These two items should be revised.

The kick contained items 38 through 41:

Item 38: "rapid continuous approach to the ball." This item was confounded by poor model fit, and thus interpretation of item parameters and item information is of no value.

Item 39: "the trunk is inclined backward during ball contact" was one of the more difficult items (b = .363). It had very good discriminatory power (a = 2.836) and provided a large amount of information (I = 2.011).

Item 40: "forward swing of the arm opposite kicking leg" best discriminated (a = 2.476) among moderate-high ability levels (b = .312). It displayed a highly peaked information curve (I = 1.533), indicating good precision in estimating moderate-high ability.

Item 41: "follow-through by hopping on the nonkicking foot" was the most difficult item on the TGMD (b = 1.133). It had high discriminatory power (a = 1.642) and provided a moderate amount of information (I = .674) over a fairly wide range of ability.

Most of the kick items measured above average ability with good psychometric properties, with the exception of item 38. Item 38 appeared to be the exceptional item of the kick because of its lack of model fit. The easiest item was item 40, followed by items 40, 39. Item 41 was the most difficult item of the TGMD.

The overhand throw was the final skill on the TGMD. It was represented by items 42 through 45:

Item 42: "a downward arc of the throwing arm initiates the windup." This was a moderately easy item (b = -.555) with very high discriminatory power (a = 2.170) and a good index of precision (I = 1.177).

Item 43: "rotation of hip and shoulder to a point where the nondominant side faces an imaginary target" had medium difficulty (*b* = .1467), very high discrimination (*a* = 2.754), and provided a large amount of information (I = 1.897).

Item 44: "weight is transferred by stepping with the foot opposite the throwing hand" had a moderate-low difficulty (*b* = 0.416), exhibited very good discrimination (*a* = 1.739) and moderate information (I = .754).

Item 45: "follow through beyond ball release diagonally across the body toward side opposite throwing arm" was one of the more difficult items on the TGMD (*b* = .571). It had very good discriminatory power and displayed a very high information curve (I = 1.286).

The overhand throw items exhibited good psychometric properties. All items fit the model, displayed very high discrimination values and information curves. Moreover, the items sampled a wide range of abilities, with item 42 as the easiest followed by items 44, 43, and 45.

Summary of IRT Item Analysis

In addition to estimating values for item parameters, the IRT item analysis provided a goodness-of-fit index and an information curve to serve as an indication of an item's precision in estimating ability. These latter indices make IRT more powerful than CTT because one may determine the independent contribution of items comprising the test. Although few items were identified as poor items with traditional item statistics, the IRT analysis identified additional items exhibiting poor psychometric properties. Many of these items require revision.

Test revision may be accomplished by replacing items that displayed poor psychometric properties or the entire test could be revised. Revision of the TGMD items may include: (a) a review the motor development and biomechanics literature to identify the major behavioral components of gross motor skill acquisition, (b) the most common components may be chosen for further analysis, (c) present the items to a large sample encompassing a wide range of abilities, (d) analyze the items with IRT methodologies, (e) replace items exhibiting poor psychometric properties with items that assess the appropriate ability with an adequate amount of precision. For the TGMD to effectively measure motor development of children from 3 to 10 years of age, one would expect the items to accurately assess a wide range of ability. For the purposes of screening, items measuring a selected mastery or ability level for each age group may be grouped and presented as a criterion-referenced test. Thus, the TGMD could effectively serve as a measure of general gross motor development and/or a screening test.

A complete test revision employs a slightly different strategy. The developer would follow the steps outlined for the revision of the TGMD items. The test developer would review the literature and choose those skills that best represent the gross motor development domain. Item response theory methodologies would be applied to evaluate the most common behavioral criteria for model fit, item difficulty, item discrimination, and precision. Depending on the nature of the test, items exhibiting appropriate characteristics should be included in the revised test. Items included in diagnostic tests should be arranged sequentially to assess a wide range of ability. Screening or mastery classification tests should include items that provide the most information at or near the mastery level cut-off score. Once the purpose of the test has been defined, and items/skills chosen, the test should be evaluated for construct and content related evidence of validity to insure that the test represents the domain of gross motor development (see Safrit, 1981 and Safrit, 1989 for further discussion of test validity).

The following skills were identified in the present study as providing insufficient precision to estimate ability accurately:

1. The run : Items 1 and 3 did not fit the two-parameter model. Item 3 measured extremely low ability,while items 3 and 4 displayed a very low information curve indicating inadequate precision for measuring ability. In addition, the locomotor subtest revealed very low precision to estimate this ability level. Thus, the TGMD should be revised to either include more items to measure low ability or the items should be replaced by items measuring a higher degree of difficulty. In either case, items 1, 3, and 4 should be revised.
2. The gallop contained two items that should be revised; (items 6 and 8). Item 6 measured low ability with low precision and item 8 did not fit the two-parameter model.
3. The horizontal jump displayed one item (18) that had poor psychometric properties. Item 18 was also identified using traditional item statistics as exhibiting poor discrimination with a low degree of difficulty. The IRT analysis confirmed the traditional findings.
4. Item 20 confounded the skip by showing a lack of fit to the two-parameter model.
5. The slide was an easy skill, but interpretation of ability is limited by the low amount of information exhibited by all items of the slide (I values ranged from .159 to .284).
6. The strike should be revised because of the low amount of information displayed by item 27 and the misfit of items 28 and 29.
7. The stationary bounce showed good psychometric properties except for item 32 which did not fit the two-parameter model and therefore should be revised.

8. The catch contained two items that should be revised. Item 34 provided limited information and item 35 displayed a poor fit of the two-parameter model.

9. The kick displayed good psychometric properties with the exception of item 38 that did not fit the two-parameter model.

The following items displayed good measurement properties: the hop, leap, and the overhand throw. The hop measured a wide range of abilities (the *b*-parameters ranged from -1.291 to .112) and displayed adequate information curves. The leap best measured moderate to moderate-high ability (*b*-parameters ranged from -.246 to .633), supplying very good information. The overhand throw also revealed moderate to moderate-high ability (*b*-parameters ranged from -.555 to .571) with very good precision.

Mastery Classification

Mastery Classification Validity of the TGMD

The purpose of mastery classification is to determine if an examinee reached a predetermined achievement or performance level. Minimum level of mastery and maximum level of nonmastery are chosen by considering the consequences of misclassifying a master or a nonmaster. Once the classification criterion has been chosen, a test is analyzed for its decision accuracy. A contingency table is drawn that compares the true state of mastery with the classification decision based on test scores. The true state of mastery may be determined by a number of methods (see Safrit, 1989 for a description of these methods), including estimation of true mastery via IRT.

Item response theory true score estimates were used in the present study to represent the true mastery state. This theory provides an estimation of examinees' latent trait and true score. It is not assumed that IRT is the absolute measure of true ability, but it provides a more precise estimation of true ability. It should be noted, however, that IRT estimation of the "true" mastery state, like other estimation methods, suffers from a degree of subjectivity in choosing the most valid cut-off score.

The present study compared IRT and CTT mastery classification using cut-off scores that represented 70% and 85% mastery. These cut-off scores were those chosen by Ulrich (1984). Ulrich analyzed the reliability of the these mastery levels for classifying handicapped and nonhandicapped children, but failed to provide empirical evidence for validity (i.e., decision accuracy) of the cut-off scores.

Raw scores representing 70% and 85% mastery for the locomotor and object control subtests were computed from the CTT statistics. The cut-off scores were calculated by multiplying 70% and 85% by the number of items on the subtests. The 70% mastery level raw scores for the locomotor and the object control subtests were 18 and 13, respectively. The 85% mastery level raw score for the locomotor subtest was 22 and 16 for the object control subtest. Those subjects exhibiting raw subtest scores at or above the criterion were classified as masters and those falling below the cut-off score were classified as nonmasters.

The computation of the IRT mastery level cut-off scores was more complex. The ability level that represented 70% and 85% mastery was calculated. Theoretically, a test characteristic curve would provide the vehicle by which a true score representing 70% and 85% mastery could be transformed into an ability score. However, the test characteristic curve was not provided by the PC-BILOG computer program. Therefore, the mastery level cut-off scores were computed utilizing the test characteristic curve equation $\Sigma P_i(\Theta)$ and the item characteristic curve equation for the two-parameter model (Baker, 1985):

$$P_i(\Theta) = 1/[1 + e^{-1(\Theta - b_i)}] \qquad (8)$$

where e = 2.718; *b* is the difficulty parameter; and (Θ) is the ability parameter.

A computer program was written to calculate the true test score for each subtest from selected ability levels, see Appendix C. Ability scores (Θ) that represented 70% mastery for the locomotor and object control subtests were (Θ) = .09 and (Θ) = .40, respectively. Ability scores that reflected 85% mastery for the locomotor and object control subtests were (Θ) = .53 and (Θ) = .89, respectively. Those scoring at or above the cut-off ability scores were classified as masters, while those whose scores fell below were classified as nonmasters.

In order to assess the agreement of mastery classification between the two test theories, four contingency tables were drawn to encompass both subtests at each mastery level cut-off score. The contingency coefficient statistic (C) reflects the proportion of subjects classified as true masters (nonmasters) using IRT as the "true" state of mastery and those classified as masters (nonmasters) by CTT. The test theories showed a high degree of agreement in classifying masters and nonmasters at the 70% and 85% mastery levels for both the locomotor and object control subtests. The locomotor subtest revealed a C of .93 and .99 for the 70% and 85% mastery levels, respectively. The object control subtest revealed higher proportions of agreement by exhibiting C-values of .99 and .997 for the 70% and 85% mastery levels, respectively. The 85% mastery levels for both subtests appeared to more accurately classify masters and nonmasters, as indicated by higher C statistics.

The kappa (*k*) statistic reflects the degree of agreement when chance agreement is considered. The *k* statistic also reflected a very high agreement in mastery classification between CTT and IRT. The *k* values for the locomotor subtests were .86 and .98 for the 70% and 85% criterion, respectively. The object control *k* statistics were computed at .94 and .99 for the 70% and 85% mastery levels, respectively. With the exception of one contingency table (70% mastery of the object control subtest), the contingency tables showed that most of the errors of classification were due to CTT classifying subjects as masters, while IRT classified the same subjects as nonmasters.

Some subjects who exhibited scores at the CTT mastery cut-off score were classified as masters under IRT, while others with the same CTT score were not classified as masters under IRT. Ability scores were calculated to four decimal places and the cut-off scores were rounded-off to two decimal places. Classical test theory employed integers for mastery classifications. Thus, subjects were classified with a greater degree of precision under IRT than CTT. Item response theory is more sensitive for classifying masters and nonmasters, since the underlying scale allows for better discrimination between masters and nonmasters.

Mastery Classification Precision of the TGMD

The second part of evaluating the validity of the mastery classification score was to determine if the TGMD provides the most information at ability levels representing the 70% and 85% mastery. Subtest information curves displayed the relative precision of the test to measure the continuum of ability. Since the subtest information curves centered around low ability levels, it is expected that the test best measures low abilities. The computer program that calculated true scores for the 70% and 85% mastery level cut-off score was also employed to determine the true scores best measured by the TGMD subtests. As expected, the locomotor subtest provided the most information at a true score of 7.7 which represents 30% mastery. The locomotor subtest was most precise at $(\Theta) = -1.857$. The object control subtest was most precise at low ability, $(\Theta) = -1.643$, which reflected a true score of 8.5 and a mastery level of 45%.

Although CTT classify masters and nonmasters accurately, as indicated by the contingency coefficients, the TGMD most precisely estimates abilities near 30% and 45% mastery for the locomotor and the object control subtests, respectively. Thus, the TGMD may be most useful for classifying individuals of low gross motor ability.

Discussion

Assessment in the psychomotor domain is challenging, especially given the current index of tests readily available. Practitioners are frustrated by the restrictions imposed by traditional test theories to assess the wide variety of student abilities found in public schools. The limitations of current tests may be in the theory that serves as the test's foundation. Most tests are based on CTT which limits the generalizations of test scores for practitioners and test designers alike.

Item response theory shows great promise in improving psychomotor measurement, especially for those who must assess atypical populations. Item response theory can be employed to develop or revise tests to meet the criteria of good tests. The item parameters are invariant across populations, thus, without severe deviations from the population, one test may be administered to measure students displaying a wide variety of characteristics. One may determine if the test is sensitive in measuring change in performance of lower abilities and tests may be constructed to assess a wide continuum of ability. A child who cannot complete or refuses to attempt an item may still be evaluated and his or her ability estimated, because IRT does not mandate that all items be administered, provided that an adequate number of items are given to insure the precision of the measure and validity of test score interpretation.

Item response theory can be employed for efficient testing sessions. With the increased availability of computers, it is not unreasonable to imagine the use of portable or lap-top computers to aid in the assessment of motor skills. This avenue of testing holds great promise for itinerant teachers and therapists. Efficient use of administration time through the use of computer technology would improve measurement in physical education.

Criterion-referenced testing and instructional programming in physical education may benefit from the application of IRT. Criteria could be arranged according to difficulty. For testing purposes a computer program could identify items or skills at the next level of difficulty. Such information may be used in instructional programming, where motor skill progressions may be based on item difficulty. Criteria based on difficulty could also be employed in the Individual Education Plan (IEP) process. A clear understanding of motor development may be incorporated into a students IEP by establishing goals, objectives, and evaluation based on a logical sequence of items. This may help establish strong educational programs.

The efficiency of diagnostic testing may be improved through the employment of IRT methodologies. Given a bank of items with known item parameters, only those items that approximate an examinee's ability need to be administered. The test may be terminated when a sufficient number of items have been given to provide a stable estimate of ability. Item response theory, coupled with other measurement methodologies such as sequential probability ratio testing (SPRT), may improve measurement in the psychomotor domain (See Safrit, Wood, Ehlert, Hooper, and Patterson (1985) for a more detailed discussion of SPRT). Sequential probability ratio testing has been shown to be an efficient model of testing motor skills (Safrit, et al., 1985), where skills are presented in a sequential manner. The minimum level of mastery and the maximum level of nonmastery are predetermined. Examinees are then administered items until their score equals one of the predetermined cut-off scores and subsequently, the examinee is classified as a nonmaster or a master. Adaptive testing procedures may be used in conjunction with SPRT models to develop efficient diagnostic tests.

Item response theory holds great promise for adapted physical educators because tests may be designed to meet the demands of the Education for All Handicapped Children's Act (EHA). Well constructed tests may be developed to assess a wide range of abilities. Test administration time may be reduced through the use of more efficient testing methods and programming and evaluation may be improved through the identification and sequential ordering of skills that lead to improved motor functioning. For the purposes of identification, diagnostic testing, programming, and evaluation, IRT may be employed at many stages of the educational process to improve services and help those who teach motor skills be accountable.

The advantages of IRT may solve many of the problems associated with measuring psychomotor skills. However, the benefits of IRT are associated with costs that may hinder the acceptance and applicability of IRT to the psychomotor domain. Some of the disadvantages to using IRT include: (a) large data sets (*n* = 200 to 1000) are required for accurate calculations of item parameters and ability estimates, (b) strong assumptions must be met to claim the advantages of IRT, such as local independence and unidimensionality, (c) item parameters and ability estimates are expressed on a scale that is difficult to interpret, (d) complex computer programs are required to obtain item parameters, ability estimates, and IRT functions (i.e., information function). Although the disadvantages of IRT may deter many researchers from applying IRT to the measurement of psychomotor skills, the results of the present study are encouraging and suggest that IRT may be utilized in improving measurement and evaluation in the psychomotor domain. The ability to analyze items independently and the characteristic of item parameter invariance makes IRT especially appealing for those interested in assessing motor skills of special populations. Although the degree to which motor skill testing can meet the assumptions of IRT is unknown, the results of this study provides evidence that the characteristics of IRT are well suited to improve measurement and evaluation in the psychomotor domain.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

Introduction

The purposes of the present study were to (a) provide insight into the use of item response theory (IRT) with psychomotor skills, (b) assess the psychometric properties of the Test of Gross Motor Development (TGMD) using IRT, and (c) provide a basis for future studies of the TGMD using IRT. Item response theory has been successfully used in the cognitive and affective domains and shows great promise for the psychomotor domain. The present study followed a recommended avenue of IRT research in the psychomotor domain (Wood, 1987). The first-stage of a five-stage plan has thus far been investigated. Research papers and discussions of the applicability of IRT to improve tests in the psychomotor domain have been put forth (Disch, 1987; Safrit, 1987; Spray, 1987,1989; Wood, 1987). The encouraging results of these papers provided the impetus for the present study and consequently the investigation of the next stage of research-examination of the current literature in the affective, cognitive, and psychomotor domains and application IRT analysis to psychomotor tests. The dichotomously scored TGMD is a test instrument which measures psychomotor skills in a cognitive test framework, thus providing a convenient "transitional" type test which can be used to examine the use of IRT with psychomotor skill tests.

Spray (1987) and Wood (1987) cautioned against the random application of IRT to psychomotor measurement without first evaluating the need for changing or improving current measurement and evaluation practices, and examining the degree to which the measurement of psychomotor skills meets the assumptions of IRT. The need to improve current methods of assessing psychomotor skills is warranted since (a) psychomotor skill assessment of special populations is mandated by the Education of

All Handicapped Children Act (EHA), and (b) literature suggests that current psychomotor skill assessment practices of children with handicapped conditions is plagued by a lack of effective tests that measure a large continuum of ability across many populations (Baumgartner & Horvat, 1988; Davis, 1984; Klesius, 1981; Seaman, 1988). And yet, the need to improve current methods of assessing psychomotor skills is warranted. However, the degree to which IRT models are robust to violations of assumptions is not well established. Research suggests that IRT models are fairly robust to violations of assumptions for knowledge and cognitive tests, but conclusions put forth are not definitive. Research regarding the application of IRT to the psychomotor domain is very limited and the degree to which assumption can be met or violated is an avenue of future research.

This study serves as the first step in a line of research aimed at improving psychomotor assessment of children with handicapping conditions. A direct consequence of this study serves as input for the revision of the TGMD. Additionally, the present study provides a basis for future research which may include assessment of test bias across various special populations, equating ability to developmental age, and implementing adaptive testing procedures.

The present study employed data used by Ulrich (1985) in his original psychometric analysis of the TGMD. The data consisted of 913 subjects aged 3 to 10 years, nonhandicapped and 20 mildly mentally handicapped. Thirty-two subjects (3.5%) were deleted from the record because they performed at 100% mastery. Item response theory cannot provide accurate estimates of ability at mastery levels of 0% and 100% because it indicates that the test is either too easy or difficult for the subjects. Consequently, the interpretation of the IRT analysis is only valid for those who exhibited a raw score of 44 to 1 on the TGMD. The TGMD data were analyzed by subtests, since a factor analysis (Ulrich, 1985) revealed that the test was not unidimensional. Analyzing the TGMD as locomotor and object control subtests was the most logical and convenient for interpretation of the results. Moreover, it was necessary so as not to violate the unidimensionality assumption of IRT.

As expected, interpretation of traditional item statistics and IRT item parameters revealed that item difficulty and item discrimination were closely related. The locomotor IRT difficulty parameters revealed a high negative correlation (r = -.87) with the CTT difficulty statistics, while the object control IRT difficulty parameters displayed a very high negative correlation (r = -.98) with their CTT counterparts. Item response theory discrimination parameters correlated highly with CTT discrimination statistics within the locomotor (r = .91) and the object control (r = .94) subtests. Thus, the two test theories put forth parallel interpretations of item characteristics.

The two-parameter logistic model provided a much better fit than the one-parameter logistic model. Only 8 items did not fit the two-parameter model, while 27 items showed a poor fit for the one-parameter model. The following findings were reported from the two-parameter IRT analysis:

1. The locomotor subtest was less difficult (median difficulty = .-944) than the object control subtest (median difficulty = .053).
2. The object control subtest displayed a better discrimination index (median = 2.17) than the locomotor subtest (median = 1.54). The object control subtest displayed very good discrimination while the locomotor subtest revealed good discriminatory power.
3. Locomotor subtest information was reported at I = 15.50, indicating adequate precision to measure low ability (Θ = -1.857). The object control information function showed that the subtest displayed more information (I = 18.24) at a slightly higher ability level (Θ = -1.643).
4. The standard error of estimation SE(Θ) was reported at .25 for the locomotor subtest and .30 for the object control subtest. The object control

subtest showed a higher margin of error in estimating lower abilities than the locomotor subtest.

5. Item analysis revealed that all items of the hop, leap, and overhand throw exhibited good psychometric properties.
6. Nine out of 12 skills contained items that displayed poor psychometric characteristics and/or did not fit the two-parameter model. The run (items 1, 3, and 4), gallop (items 6 and 8), horizontal jump (item 18), skip (item 20), slide (items 23, 24, 25, and 26) , strike (items 27, 28, and 29), stationary bounce (item 32), catch (items 34 and 35), and the kick (item 38) should be revised.
7. The cut-off score analysis of IRT and CTT showed a high agreement of classifying masters and nonmasters at the 70% and 85% levels of mastery. The locomotor subtest revealed a decision validity coefficient of .93 and .99 for the 70% and 85% mastery levels, respectively. The object control subtest revealed higher decision validity coefficients of .99 and .997 for the 70% and 85% mastery levels, respectively.
8. The TGMD subtests were found to best measure very low mastery levels, where the most precision for measuring ability represented 30% and 45% mastery for the locomotor and object control subtests, respectively.

Conclusions

Based on the findings of this study, the following conclusions are warranted:

1. Item response theory using the two-parameter logistic model provides an effective psychometric analysis of dichotomously scored psychomotor skills.

2. Item response theory addresses many of the shortcomings of CTT, such as the inability to generalize item statistics to various populations and to determine the contribution of test items independently. The invariance property of IRT is very appealing to those who must assess atypical populations because a single test can accommodate various populations and wide ranges of ability.
3. Item response theory provides a more in depth understanding of the psychometric properties of the TGMD. In addition to item difficulty and discrimination, the precision of each item and subtests to measure a given ability level can be determined.

Recommendations

The following are recommended for future studies involving IRT and test development in the psychomotor domain:

1. Items of the TGMD displaying poor psychometric properties should be revised.
2. The TGMD should have a hierarchical arrangement of items to facilitate test administration and motivation for the examinees requires examination via item information functions.
3. The use of IRT in the development of adaptive testing strategies to improve the efficiency of gross motor development assessment. Adaptive testing could also be coupled with other measurement techniques, such as sequential probability ratio testing, to improve measurement of motor skills.

4. The use of IRT to investigate item and test bias in the TGMD to determine if the item parameters are equivalent for special populations, such as populations with different levels of mental retardation.
5. PC-BILOG did not compute accurate standard errors of item parameters and test characteristic curves. Standard errors of item parameters are necessary for item banking and adaptive testing, while test characteristic curves are useful in determining the true score that represents ability in determining mastery level cut-off scores and obtaining the true score best measured by the test. True scores are also easier to interpret than ability estimates and may be helpful for researchers and practitioners. Computer programs such as LOGIST (Wingersky, 1983), BICAL (Wright & Stone, 1979), or MULTILOG (Thissen, 1983) should be evaluated for their utility in providing such computations.
6. Future investigations are needed to assess the robustness of the IRT assumption with psychomotor skills. The interpretations of an analysis is only as valid as the degree to which the underlying assumptions have been met. For example, test designers must give specific directions and well described items to meet the IRT local independence assumption. Every effort should be made to insure a consistent testing environment.
8. Further investigation of the usefulness of one-parameter models in the assessment of psychomotor tests. The one-parameter is easier for item banking and equating, but might not be the most accurate for assessing motor skills.

REFERENCES

Baker, F. B. (1985). The basics of item response theory. Portsmouth, NH: Heinemann.

Baumgartner, T. A., & Horvat, M. A. (1988). The problems for measuring the physical and motor performance of the handicapped. Journal of Physical Education.Recreation. and Dance. 59(1), 48-52.

Birnbaum, A. (1968). Some latient trait models and their use in inferring an examinee's ability. In F. M. Lord & M.R. Novick, Statistical theories of mental test scores (pp. 453-479). Reading MA: Addison-Wesley

Bock, D. R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM alogarithm. Psychometrika. 46. 443-459.

Bruininks, R. H. (1978). The Bruininks-Oseretsky test of motor proficiency. Circle Pines, MN: American Guidance Service.

Charoenruk, K. (1989). The application of item response theory in the cross-cultural validation of the physical estimation and attraction scale. Unpublished doctoral dissertation, Oregon State University.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R.L. Linn (Ed.), Educational Measurement (pp. 201-219). New York, NY: American Council on Education and Macmillan Publishing.

Costa, G. M. (1986). Application of item response theory to a motor skill test in physical education. Unpublished doctoral dissertation, University of Wisconsin at Madison.

Cratty, B. J. (1986). Perceptual and motor development in infants and children. Englewood Cliffs, NJ: Prentice-Hall.

Davis, W. E. (1984). Motor ability assessment of populations with handicapping conditions: Challenging basic assumptions. Adapted Physical Activity Quarterly. 1. 125-137.

deGruijter, D. M. N., & Hambleton, R. K. (1983). Using item response models in criterion-referenced test item selection. In R. K. Hambleton (Ed.), Applications of item response theory (pp.142-154). Vancouver, BC: Educational Research Institute of British Columbia.

Disch, J. (1987). Recent developments in measurement and possible applications to the measurement of psychomotor behavior: A response. Research Quarterly for Exercise and Sport. 58. 210-212.

Disch, J (1989). Selected multivariate statistical techniques. In M.J. Safrit & T.M. Wood (Eds.), Measurement concepts in physical education and exercise science (pp.155-177). Champaign Il: Human Kinetics.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational Measurement (pp.147-200). New York, NY: American Council on Education and Macmillan Publishing.

Hambleton, R. K., & Cook, L. L. (1977). A word about the issue. Journal of Educational Measurement. 14. 75-96.

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision on ability estimates. In D. J. Weiss (Ed.), New horizons in testing: Latient trait theory and computerized adaptive testing (pp. 31-49) New York, NY: Academic Press.

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

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

Kane, M. (1987). On the use of IRT models with judgmental standard setting procedures. Journal of Educational Measurement, 24, 333-345.

Klesius, S. E. (1981). Measurement and Evaluation: The neglected element in physical education for the handicapped. Physical Educator, 38, 15-19.

Langendorfer, S. (1986). [Review of the Test of Gross Motor Development]. Adapted Physical Activity Quarterly, 3, 186-190.

Lord, F. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 118-138.

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

Loyd, B. H. (1988). Implications of item response theory for the measurement practitioner. Journal of Applied Measurement in Education, 1, 135-143.

MacDonald, R. P. (1982). Linear versus non-linear models in item response theory. Applied Psychological Measurement, 6, 379-396.

Magill, R. A. (1985). Motor learning, Concepts and applications. Dubuque, IA: Wm. C. Brown

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 14, 139-160.

Mislevy, R. J., & Bock, D. B. (1986). PC-BILOG: Item response analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Morrow, J. R. (1989). Generalizability theory. In M. J. Safrit & T. M. Wood (Eds.), Measurement concepts in physical education and exercise science (pp.73-96). Champaign Il: Human Kinetics.

Rasch, G. (1960). Probablistic models for some intelligent and attainment tests. Copenhagen Denmark: Danish Institute for Educational Research.

Safrit, M. J. (1981). Evaluation in physical education. Englewood Cliffs, NJ: Prentice Hall.

Safrit, M. J. (1987). The applicability of item response theory to tests of motor behavior. Research Quarterly for Exercise and Sport, 58, 213-215.

Safrit, M. J. (1989). Criterion-referenced measurement: Validity. In M. J. Safrit & T. M. Wood (Eds.), Measurement concepts in physical education and exercise science (pp.119-136). Champaign Il: Human Kinetics.

Safrit, M. J., Wood, T. M., Ehlert, S. A., Hooper, L.M., & Patterson, P. (1985). The application of sequential probability ratio testing to a test of motor skill. Research Quarterly for Exercise and Sport, 56, 58-65.

Seaman, J. A. (1988). The challenge. Journal of Physical Education, Recreation, and Dance, 59, 32-33.

Shannon, G. A., & Cliver, B. A. (1987). An application of item response theory in the comparison of four conventional item discrimination indices for criterion-referenced tests. Journal of Educational Measurement, 24, 347-356.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. Review of Educational Research, 56, 495-529.

Spray, J. (1987). Recent developments in measurement and possible applications to the measurement of psychomotor behavior. Research Quarterly for Exercise and Sport, 58, 203-209.

Spray, J. (1989). New approaches to solving measurement problems. In M.J. Safrit & T.M. Wood (Eds.), Measurement concepts in physical education and exercise science (pp.229-248). Champaign Il: Human Kinetics.

Steffens, K. M., Semmes, R., Werder, J. K., & Bruininks, R. H. (1987). Relationship between quantitative and qualitative measures of motor development. Perceptual and Motor Skills, 64, 985-986.

Thissen, D. M. (1983). MULTILOG: Item analysis and scoring with multiple category response models. Chicago: International Educational Services.

Ulrich, D. A. (1984). The reliability of classification decisions made with the objectives-based motor skill assessment instrument. Adapted Physical Activity Quarterly, 1, 52-60.

Ulrich, D. A. (1985). Test of Gross Motor Development. Austin, TX: ProEd.

Ulrich, D. A., & Ulrich, B. D. (1984). The objectives-based motor skill assessment instrument: Validation of instructional sensitivity. Perceptual Motor Skills, 59, 175-179.

Ulrich, D. A., & Wise, S. L. (1984). The reliability of scores obtained with the objectives-based motor skill assessment instrument. Adapted Physical Activity Quarterly, 1, 230-239.

Werder, J. K., & Bruininks, R. H. (1987). Relationship between quantitative and qualitative measures of motor development. Perceptual Motor Skills, 64, 985-986.

Werder, J. K., & Bruininks, R. H. (1988). Body Skills: A motor development curriculum for children. Circle Pines, MN: American Guidance Service.

Werder J. K., & Kalakian, L. H. (1985). Assessment in adapted physical education. Minneapolis, MN: Burgess Publishing.

Williams, H. G. (1983). Perceptual and motor development. Englewood Cliffs, NJ: Prentice Hall.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R.K. Hambleton (Ed.), Applications of item response theory (pp. 45-56). Vancourver, BC: Educational Research Institute of British Columbia.

Wood, T. M. (1987). Putting item response theory into perspective. Research Quarterly for Exercise and Sport, 58, 216-220.

Wood, T. M. (1989). The changing nature of norm-referenced validity. In M. J. Safrit & T.M. Wood (Eds.), Measurement concepts in physical education and exercise science (pp.23-44). Champaign Il: Human Kinetics.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press

Yen, W. M. (1983). The choice of scale for educational measurement: An item response theory perspective. Journal of Educational Measurement, 20, 299-325.

APPENDICES

APPENDIX A

Items Comprising the Test of Gross Motor Development

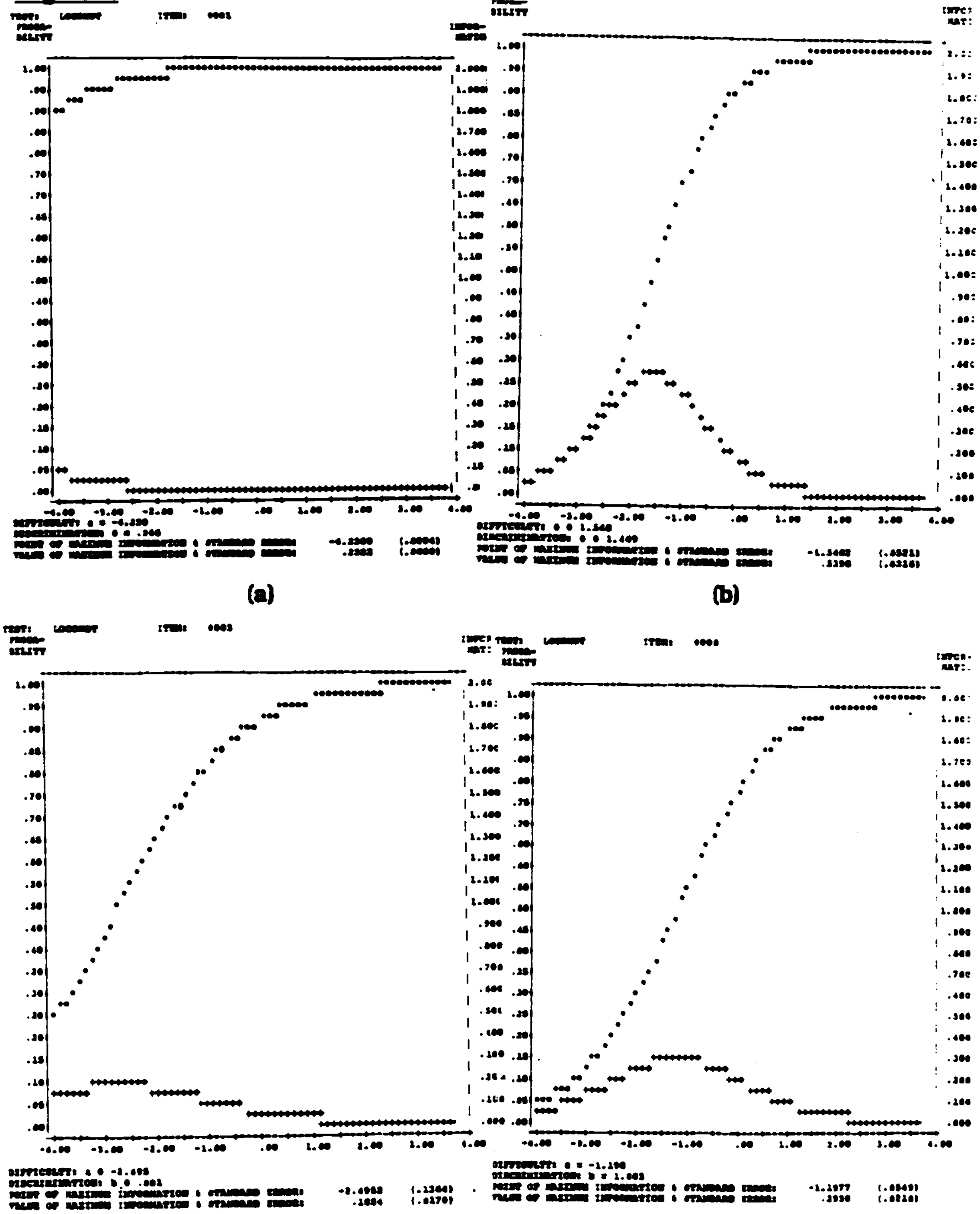
Skill	Item	Definition
Run	1	Brief period where both feet are off the ground.
	2	Arms in opposition to legs, elbows bent.
	3	Nonsupport leg bent approximately near or on the line (not flat footed).
	4	Nonsupport leg bent approximately 90 degrees (close to buttocks).
Gallop	5	A step forward with the lead foot followed by a step with the trailing foot to a position adjacent to or behind the lead foot.
	6	Brief period where both feet are off the ground.
	7	Arms bent and lifted with the right and left foot.
	8	Able to hop on the right and left foot.
Hop	9	Foot of nonsupport leg is bent and carried in back of the body.
	10	Nonsupport leg swing in pendular fashion to produce force.
	11	Arms bent at elbows and swing forward on take off.
	12	Able to hop on the right and left foot.
Leap	13	Take off on one foot and land on the opposite foot
	14	A period where both feet are off the ground (longer than running).
	15	Forward reach with arm opposite the lead foot.
Horizontal Jump	16	Preparatory movement includes flexion of both knees with arms extended behind the body.
	17	Arms extend forcefully upward, reaching full extension above the head.
	18	Take off and land on both feet simultaneously.
	19	Arms are brought downward during the landing.
Skip	20	A rhythmical repetition of the step-hop .
	21	Foot of nonsupport leg is carried near surface during hop.

	22	Arms alternately moving in opposition to legs at about waist level.
Slide	23	Body turned sideways to desired direction of travel.
	24	A step sideways followed by a slide of the trailing foot to a point next to the lead foot.
	25	A short period where both feet are off the floor.
	26	Able to slide to the right and to the left side.
Two-Handed Strike	27	Dominant hand grips bat above nondominant hand.
	28	Nondominant side of body faces tosser (feet parallel).
	29	Hip and spine rotation.
	30	Weight is transferred by stepping with front foot.
Stationary Bounce	31	Contacts ball with one hand at about hip height.
	32	Pushes ball with fingers (not a slap).
	33	Ball contacts floor on the side of the hand being used.
Catch	34	Preparation phase where elbows are flexed and hands are in front of body.
	35	Arms extend in preparation for ball contact.
	36	Ball is caught and controlled by hands only.
	37	Elbows bend to absorb force.
Kick	38	Rapid continuous approach to the ball.
	39	The trunk is inclined backward during ball contact.
	40	Forward swing of the arm opposite kicking leg.
	41	Follow-through by hopping on the nonkicking foot.
Overhand Throw	42	A downward arc of the throwing arm initiates the windup.
	43	Rotation of hip and shoulder to a point where the nondominant side faces an imaginary target.
	44	Weight is transferred by stepping with the foot opposite the throwing hand.
	45	Follow-through beyond ball release diagonally across body toward side opposite throwing arm.

APPENDIX B

Item Information and Charateristic Curves

Figure 1. Item Information and Characteristic Curves for the Run



(a)

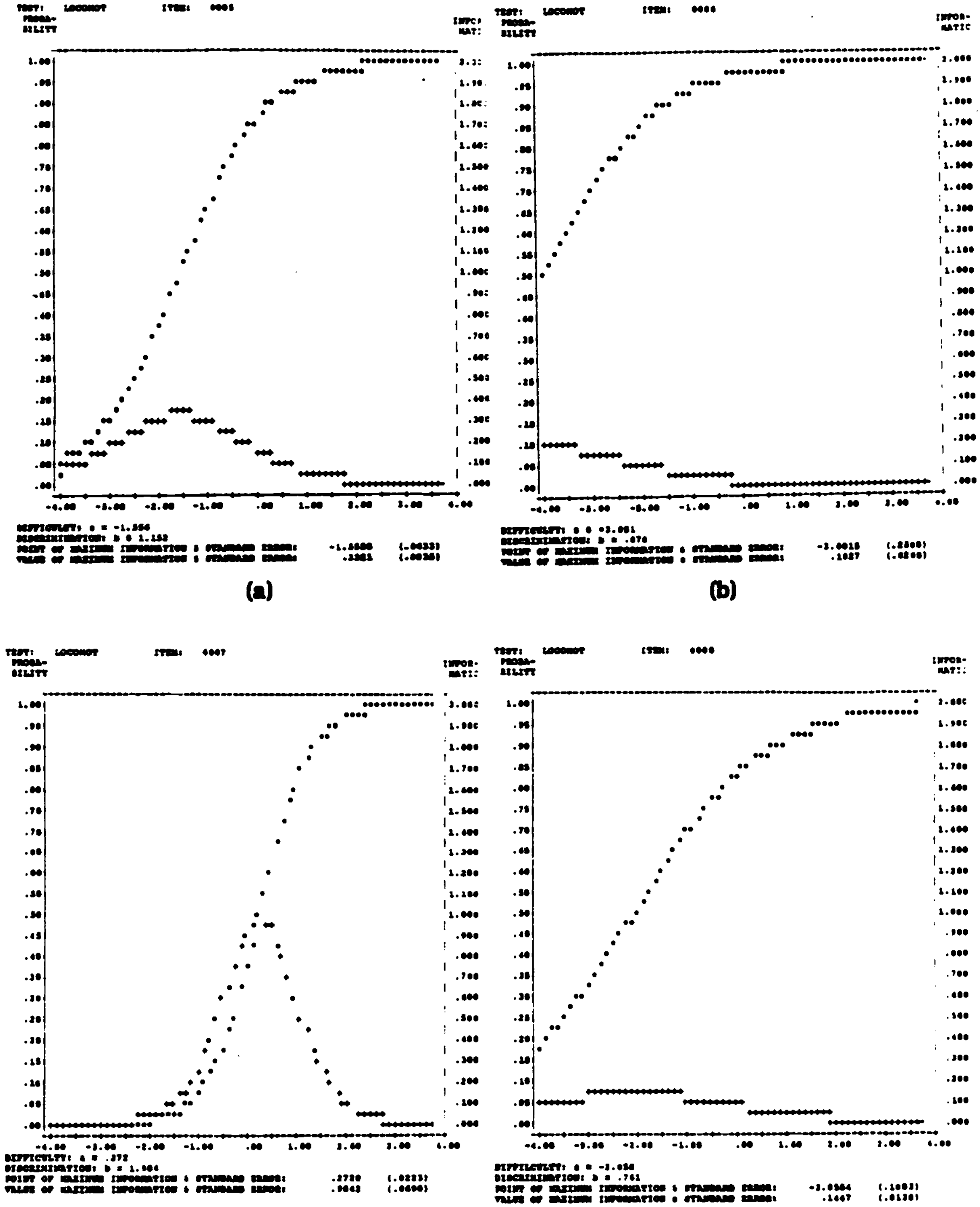
(b)

(c)

(d)

Appendix Figure 2.

Item Information and Characteristic Curves for the Gallop



```
DIFFICULTY: a = -1.356
DISCRIMINATION: b = 1.132
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:     -1.3586    (.0633)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      .3381    (.0036)
```

(a)

```
DIFFICULTY: a = -3.061
DISCRIMINATION: b = .676
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:     -3.0018    (.2508)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      .1027    (.0208)
```

(b)

```
DIFFICULTY: a = .272
DISCRIMINATION: b = 1.964
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      .2720    (.0223)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      .9642    (.0496)
```

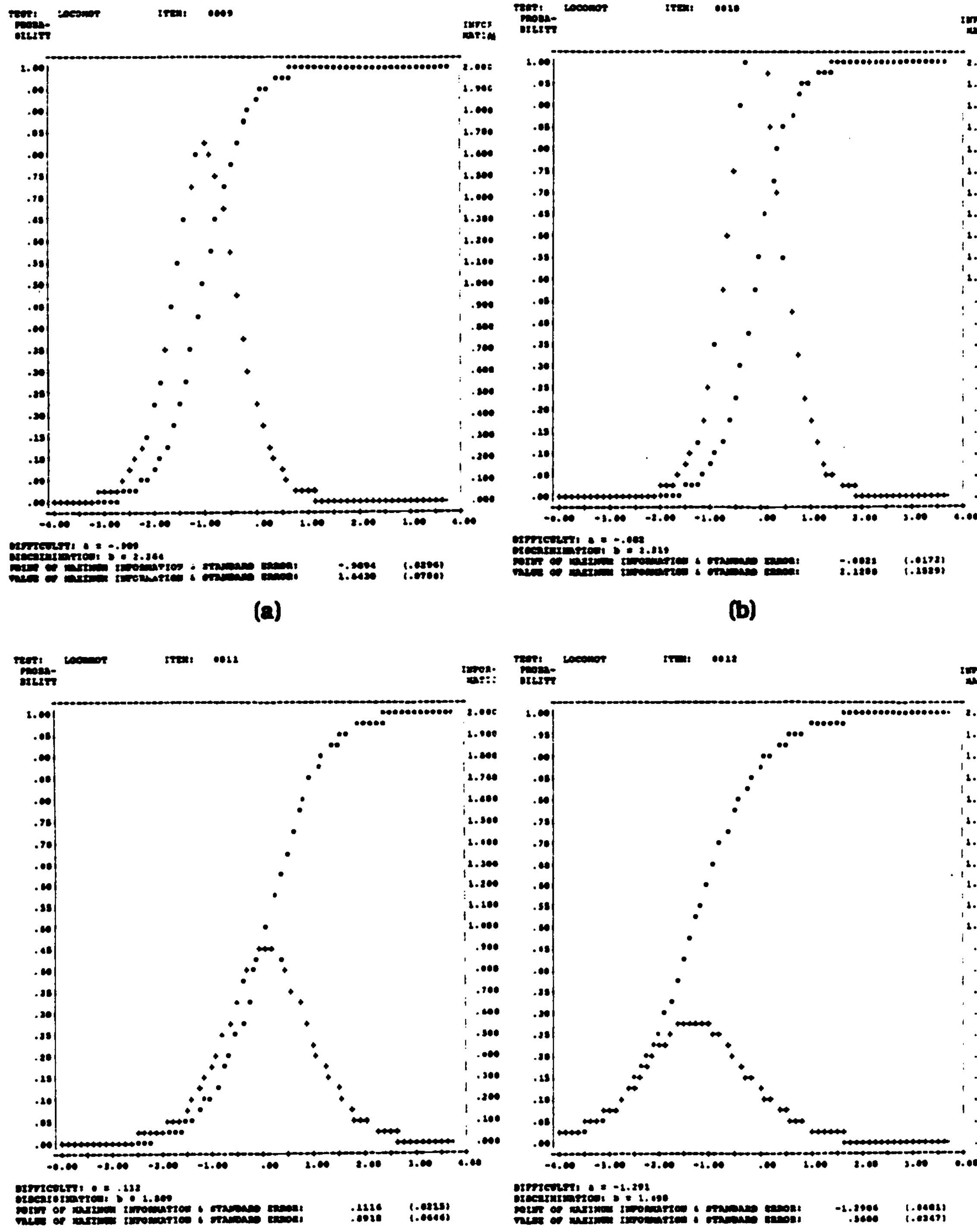
(c)

```
DIFFICULTY: a = -2.036
DISCRIMINATION: b = .761
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:     -2.0364    (.1003)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      .1447    (.0130)
```

(d)

Appendix Figure 3.

Item Information and Characteristic Curves for the Hop



```
DIFFICULTY: a = -.909
DISCRIMINATION: b = 2.344
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      -.9094    (.0296)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      1.6430    (.0780)
```

(a)

```
DIFFICULTY: a = -.002
DISCRIMINATION: b = 2.319
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      -.0021    (.0172)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      2.1200    (.1829)
```

(b)

```
DIFFICULTY: a = .112
DISCRIMINATION: b = 1.809
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:       .1116    (.0215)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .8918    (.0646)
```

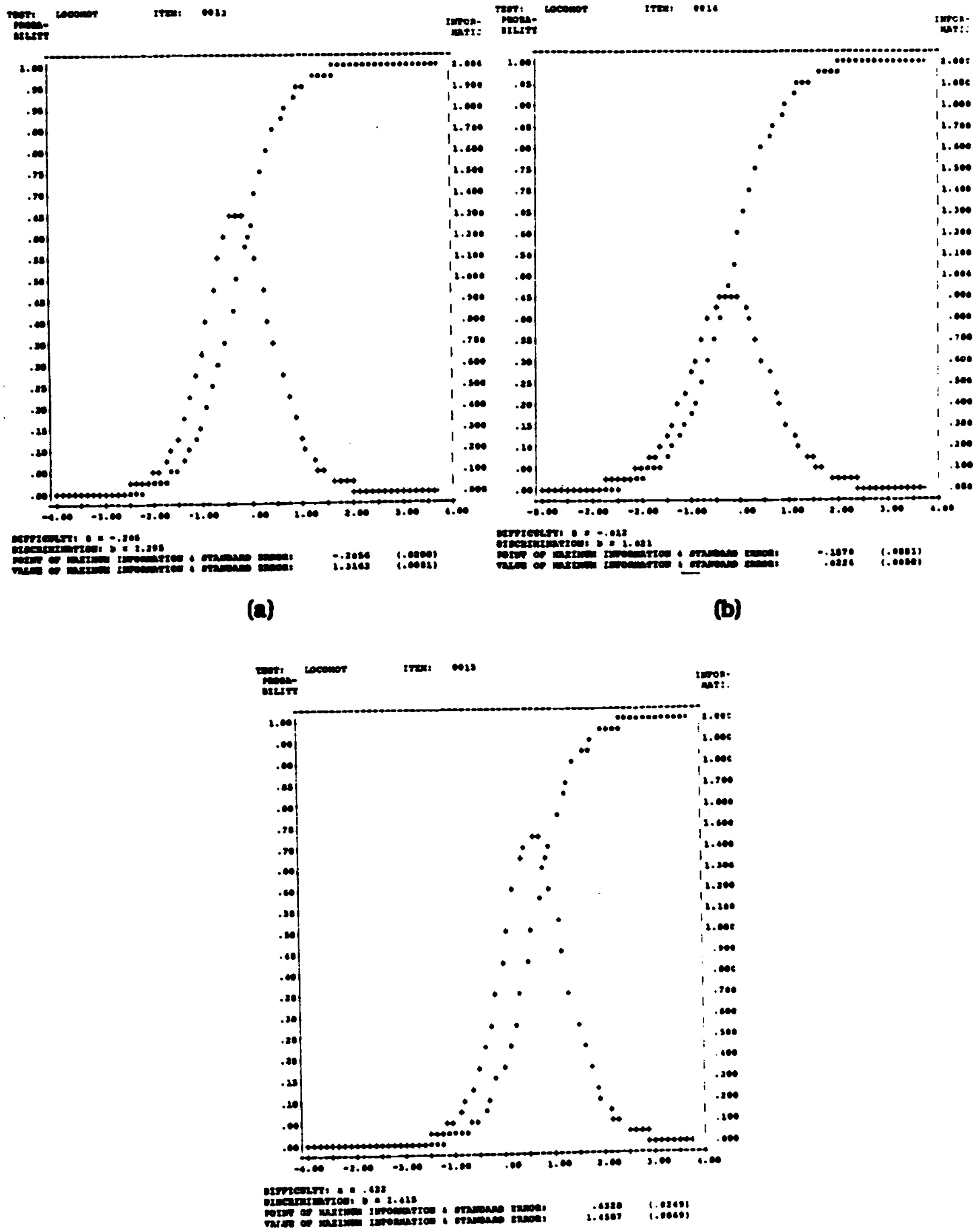
(c)

```
DIFFICULTY: a = -1.291
DISCRIMINATION: b = 1.490
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:     -1.2906    (.0401)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .5600    (.0247)
```

(d)

Appendix Figure 4.

Item Information and Characteristic Curves for the Leap



TEST: LOCOMOT ITEM: 0012

DIFFICULTY: a = -.206
DISCRIMINATION: b = 1.295
POINT OF MAXIMUM INFORMATION & STANDARD ERROR: -.2056 (.0200)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR: 1.3163 (.0081)

(a)

TEST: LOCOMOT ITEM: 0014

DIFFICULTY: a = -.012
DISCRIMINATION: b = 1.021
POINT OF MAXIMUM INFORMATION & STANDARD ERROR: -.1870 (.0081)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR: .6224 (.0050)

(b)

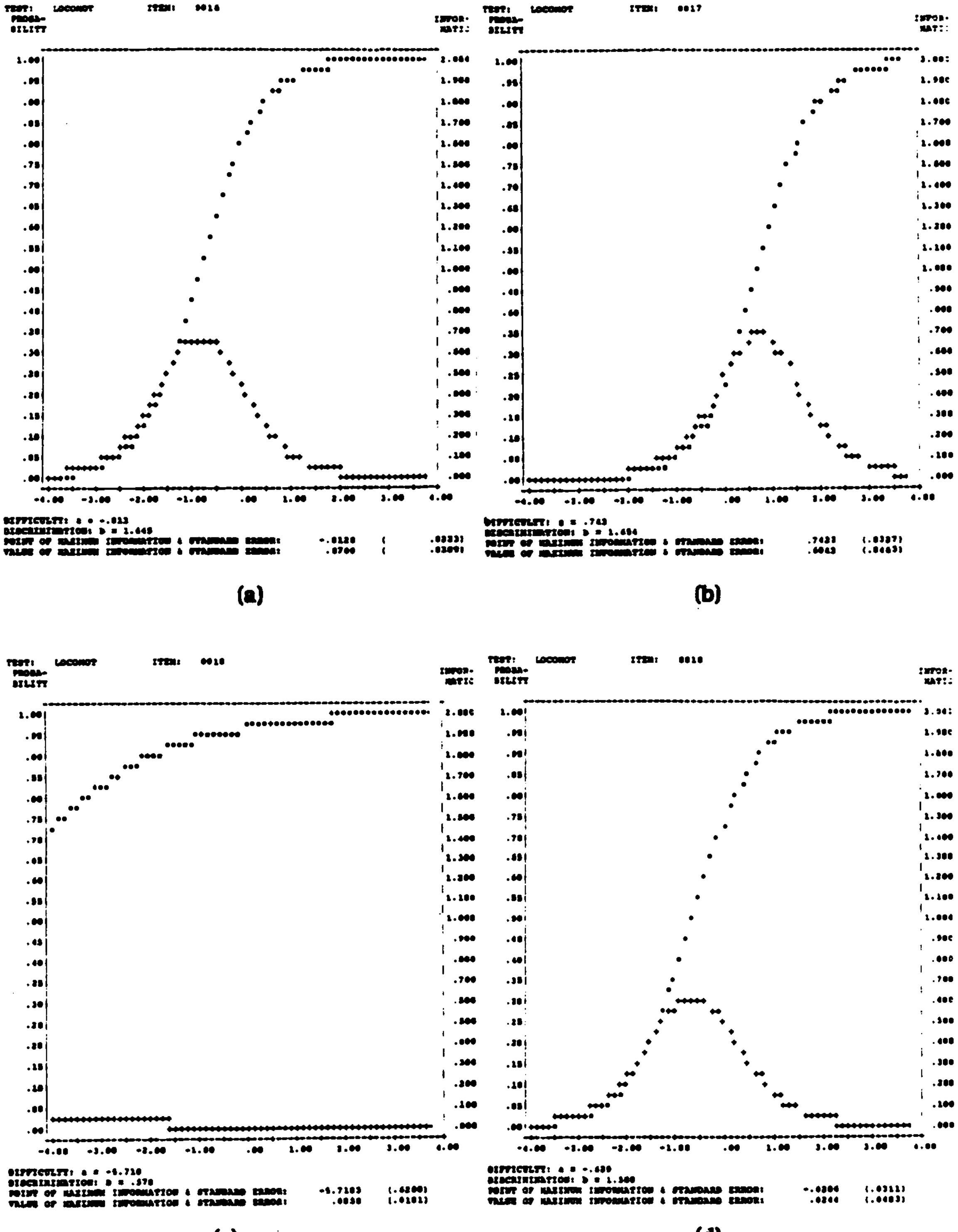
TEST: LOCOMOT ITEM: 0013

DIFFICULTY: a = .432
DISCRIMINATION: b = 1.415
POINT OF MAXIMUM INFORMATION & STANDARD ERROR: .4328 (.0249)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR: 1.4587 (.0069)

(c)

Appendix Figure 5.

Item Information and Characteristic Curves for the Horizontal Jump



```
DIFFICULTY: a = -.012
DISCRIMINATION: b = 1.448
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      -.0120   (    .0333)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .0700   (    .0389)
```

(a)

```
DIFFICULTY: a = .743
DISCRIMINATION: b = 1.494
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:       .7433   (.0387)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .6043   (.0463)
```

(b)

```
DIFFICULTY: a = -4.710
DISCRIMINATION: b = .378
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:     -4.7103   (.6200)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .0838   (.0101)
```

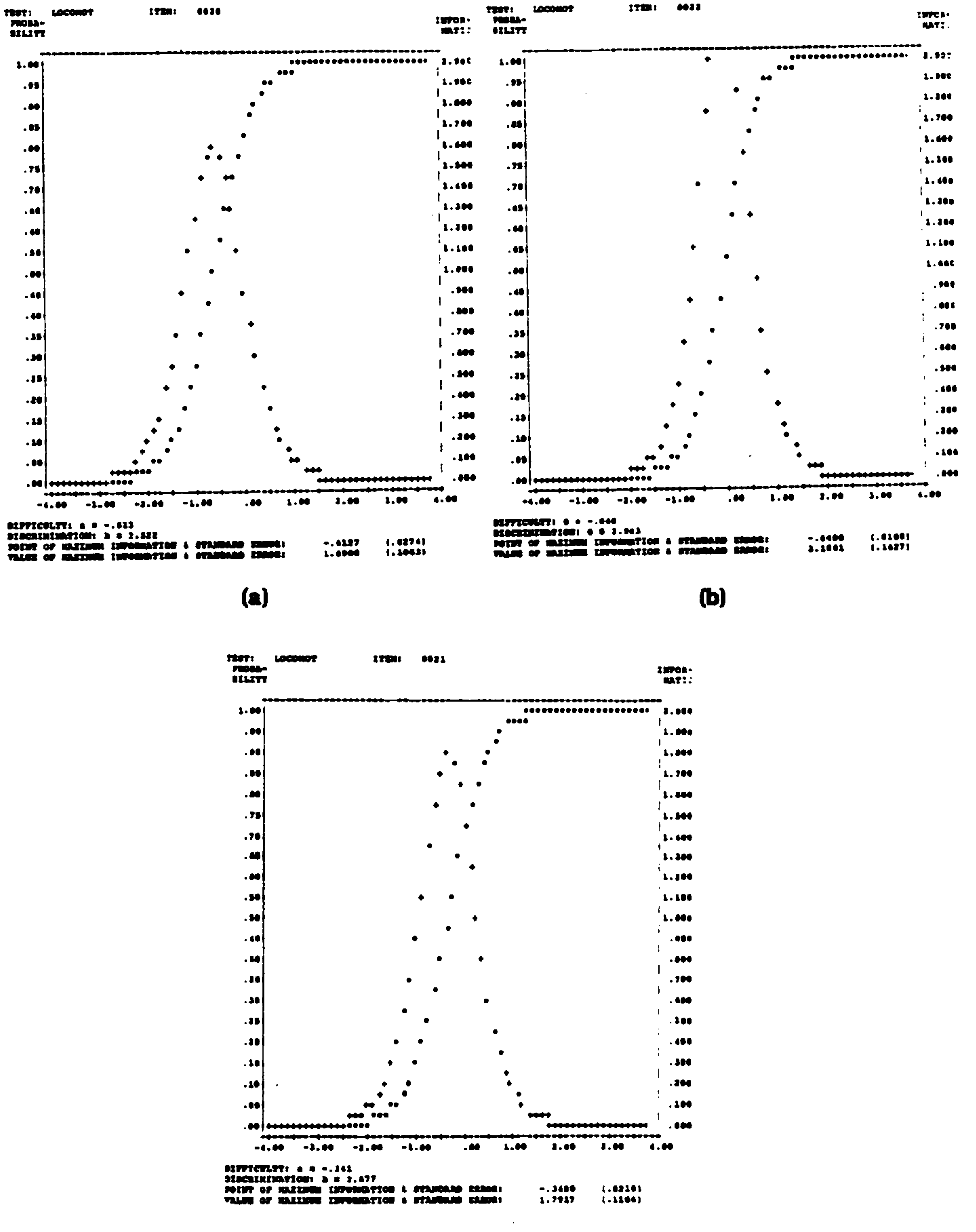
(c)

```
DIFFICULTY: a = -.030
DISCRIMINATION: b = 1.300
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      -.0304   (.0311)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .0244   (.0483)
```

(d)

Appendix Figure 6.

Item Information and Characteristic Curves for the Skip



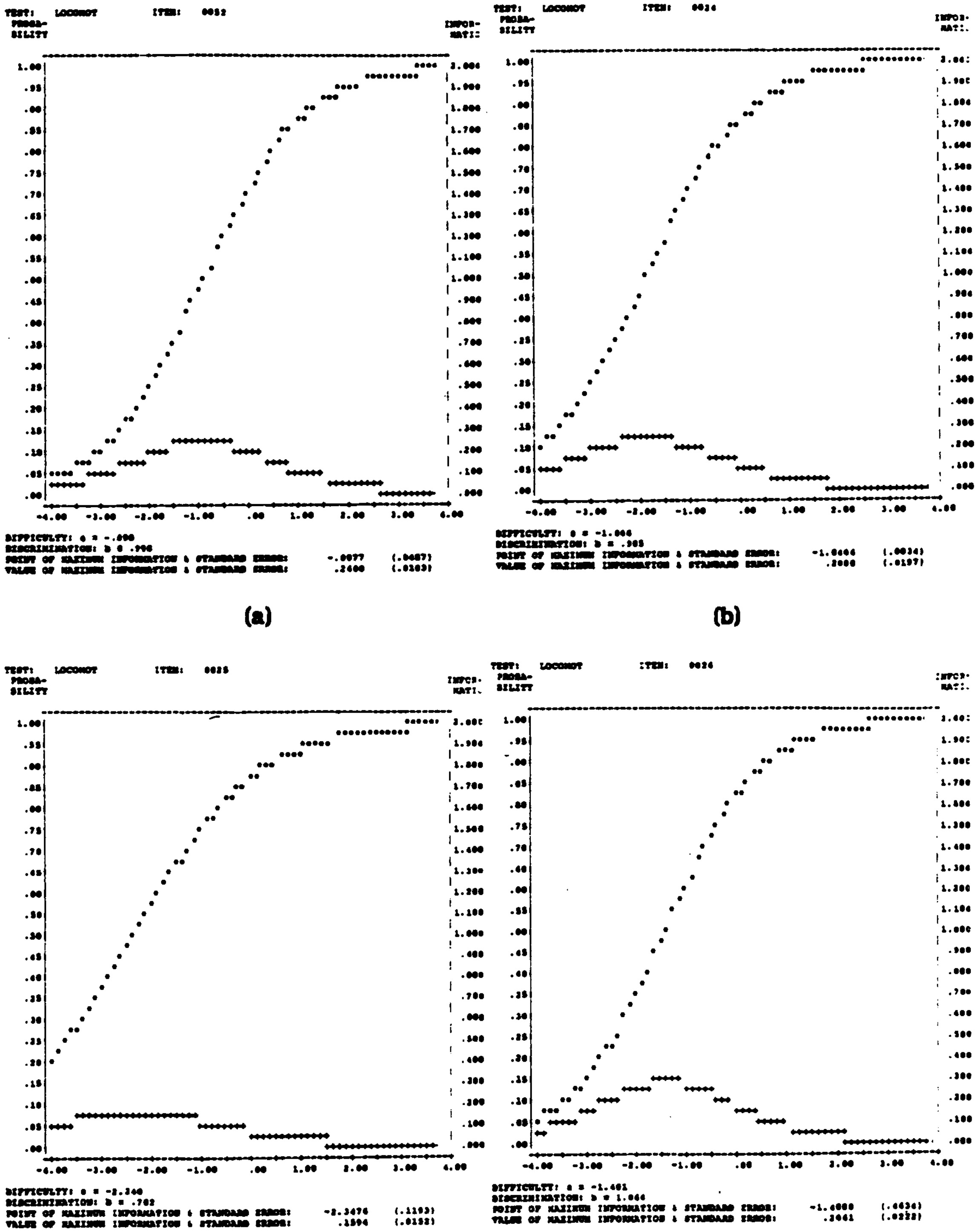
(a)

(b)

(c)

Appendix Figure 7.

Item Information and Characteristic Curves for the Slide



```
TEST:   LOCOMOT          ITEM:   0032
DIFFICULTY: a = -.090
DISCRIMINATION: b = .796
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      -.0077    (.0487)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .2400    (.0203)
```

(a)

```
TEST:   LOCOMOT          ITEM:   0034
DIFFICULTY: a = -1.066
DISCRIMINATION: b = .905
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      -1.0404    (.0034)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:        .2000    (.0197)
```

(b)

```
TEST:   LOCOMOT          ITEM:   0035
DIFFICULTY: a = -2.340
DISCRIMINATION: b = .702
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      -2.3474    (.1193)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:        .1594    (.0152)
```

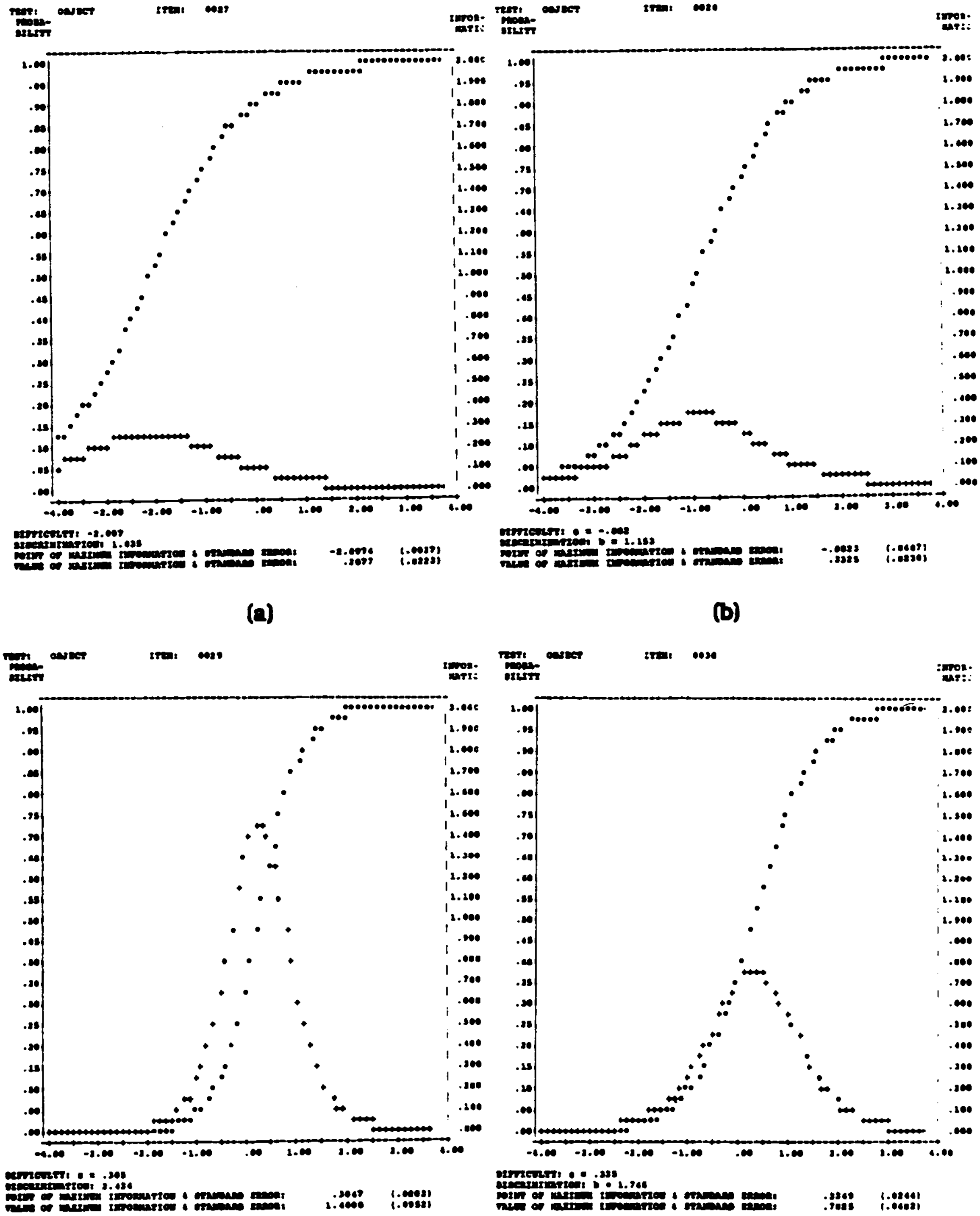
(c)

```
TEST:   LOCOMOT          ITEM:   0036
DIFFICULTY: a = -1.481
DISCRIMINATION: b = 1.044
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      -1.4000    (.4634)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:        .2461    (.0222)
```

(d)

Appendix Figure 8.

Item Information and Characteristic Curves for the Two-Hand Strike



(a)

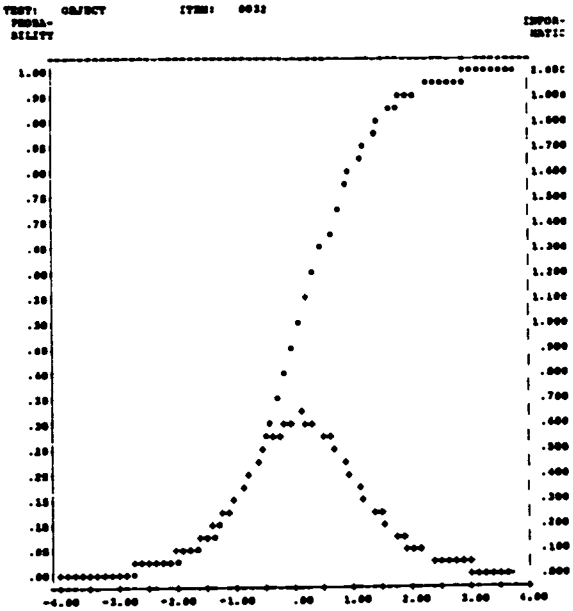
(b)

(c)

(d)

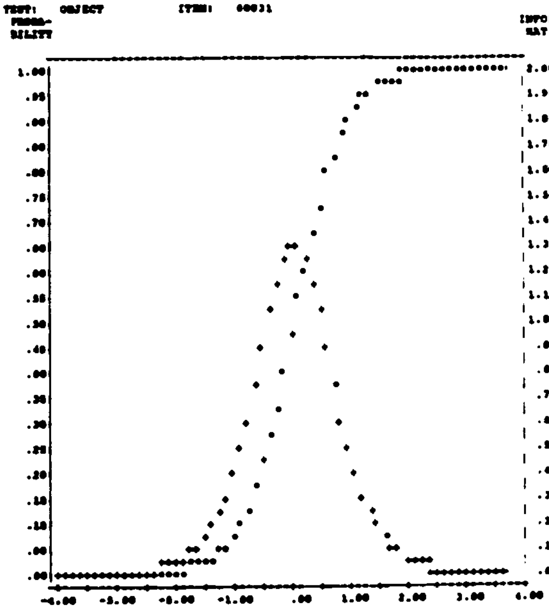
Appendix Figure 9.

Item Information and Characteristic Curves for the Stationary Bounce



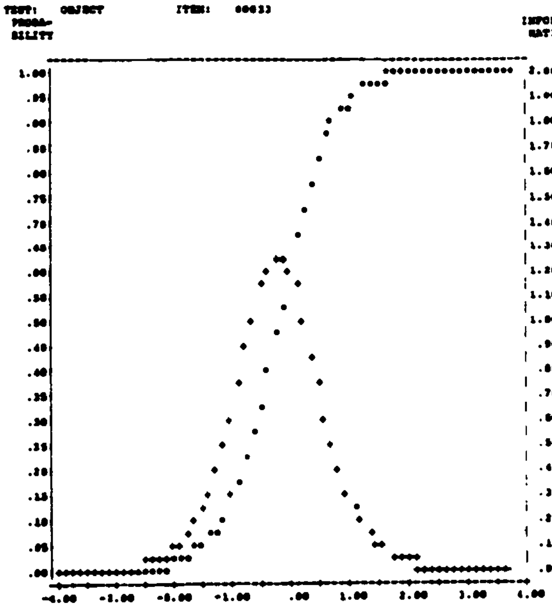
```
DIFFICULTY: a = .120
DISCRIMINATION: b = 1.183
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      .1277    (.0245)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      .6263    (.0413)
```

(a)



```
DIFFICULTY: a = .083
DISCRIMINATION: b = 2.300
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      .0828    (.0195)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:     1.3330    (.0067)
```

(b)

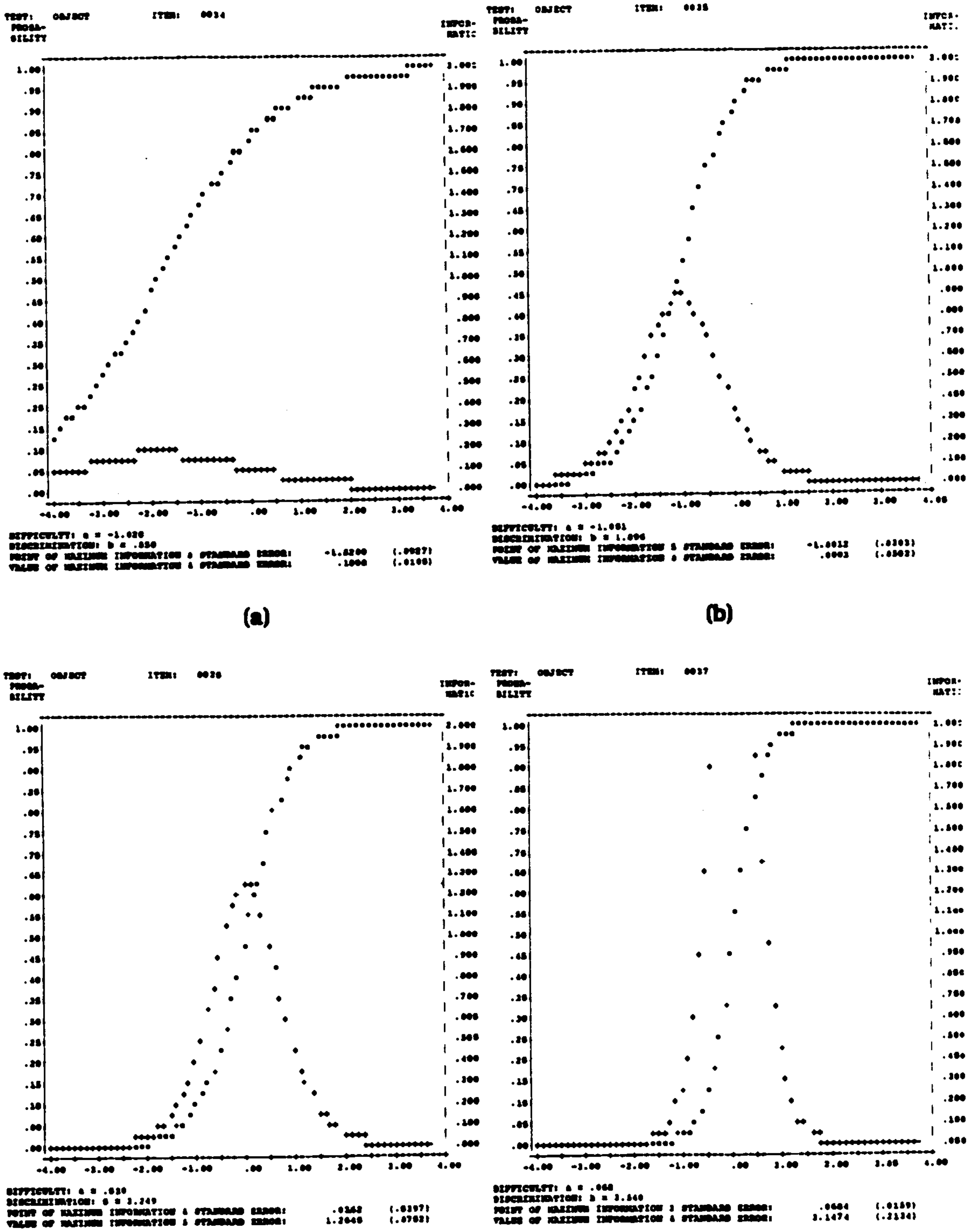


```
DIFFICULTY: a = -101
DISCRIMINATIN: b = 2.201
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:     -.1906    (.0280)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:     1.2778    (.0071)
```

(c)

Appendix Figure 10.

Item Information and Characteristic Curves for the Catch



```
DIFFICULTY: a = -1.026
DISCRIMINATION: b = .830
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:     -1.8200    (.0927)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .1000    (.0108)
```

(a)

```
DIFFICULTY: a = -1.081
DISCRIMINATION: b = 1.096
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:     -1.0012    (.0303)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .0003    (.0302)
```

(b)

```
DIFFICULTY: a = .030
DISCRIMINATION: b = 1.249
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:       .0342    (.0397)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      1.2048    (.0782)
```

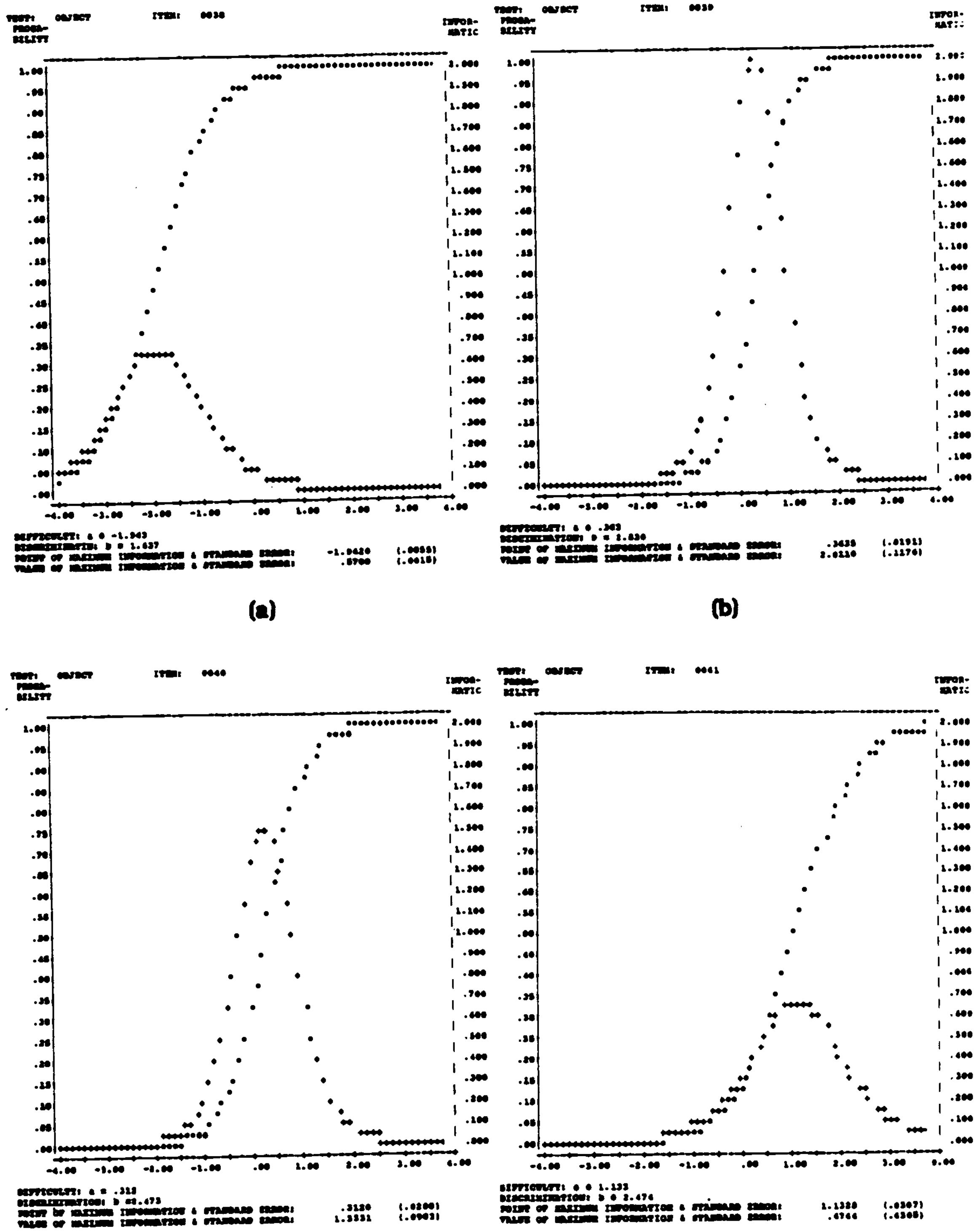
(c)

```
DIFFICULTY: a = .068
DISCRIMINATION: b = 3.340
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:       .0684    (.0159)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      3.1474    (.2134)
```

(d)

Appendix Figure 11.

Item Information and Characteristic Curves for the Kick



```
DIFFICULTY: a = -1.943
DISCRIMINATION: b = 1.637
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      -1.9420     (.0055)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:        .5700     (.0015)
```

(a)

```
DIFFICULTY: a = .363
DISCRIMINATION: b = 2.830
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:        .3635     (.0191)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       2.0110     (.1170)
```

(b)

```
DIFFICULTY: a = .312
DISCRIMINATION: b =2.473
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:        .3120     (.0200)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       1.3331     (.0902)
```

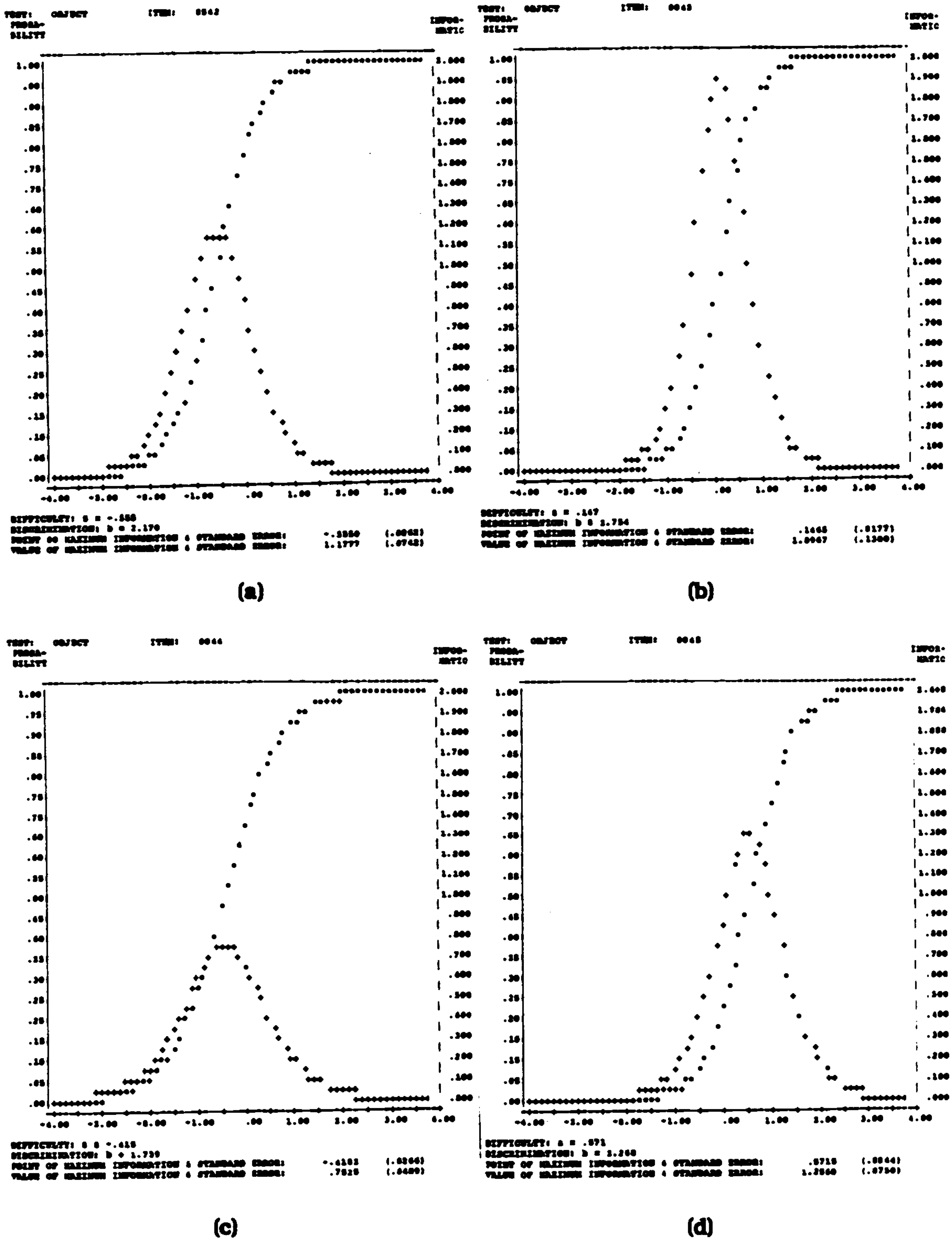
(c)

```
DIFFICULTY: a = 1.133
DISCRIMINATION: b = 2.474
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:       1.1320     (.0307)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:        .4744     (.0305)
```

(d)

Appendix Figure 12.

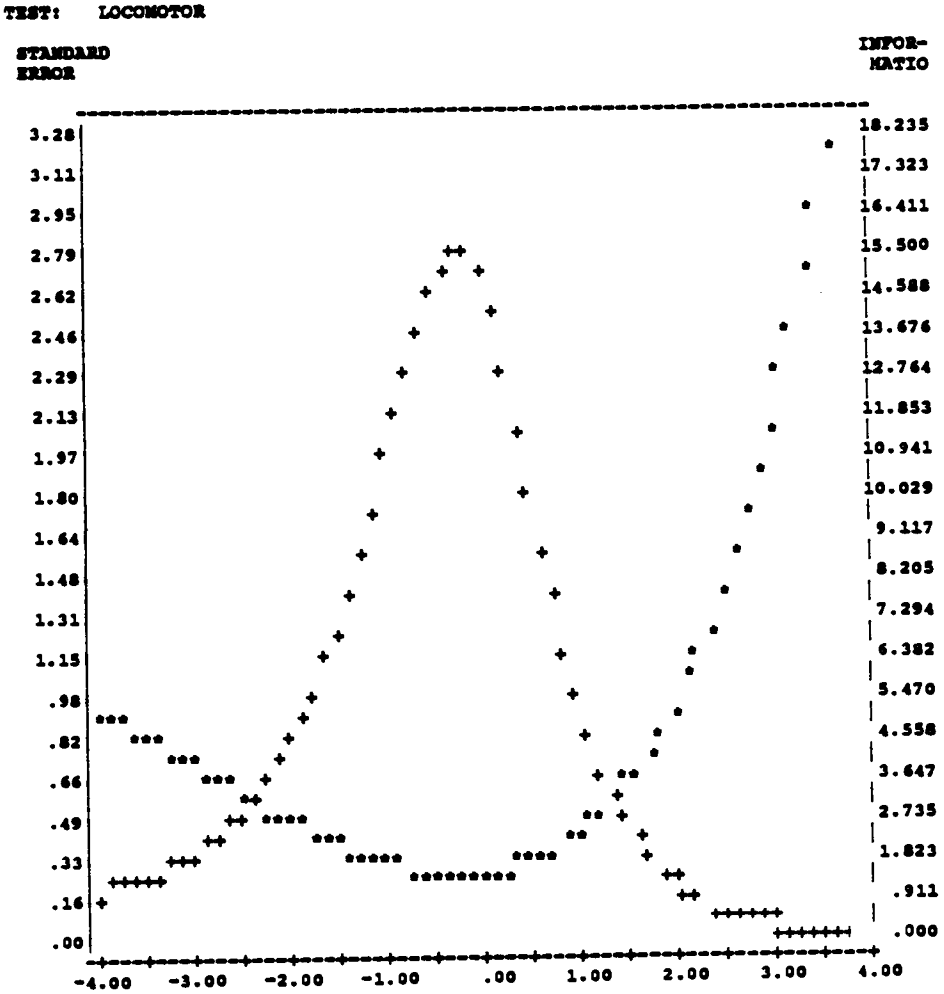
Item Information and Characteristic Curves for the Overhand Throw



APPENDIX C
Subtest Information and Standard Error Curves

Appendix Figure 13.

Locomotor Subtest Information and Standard Error Curve

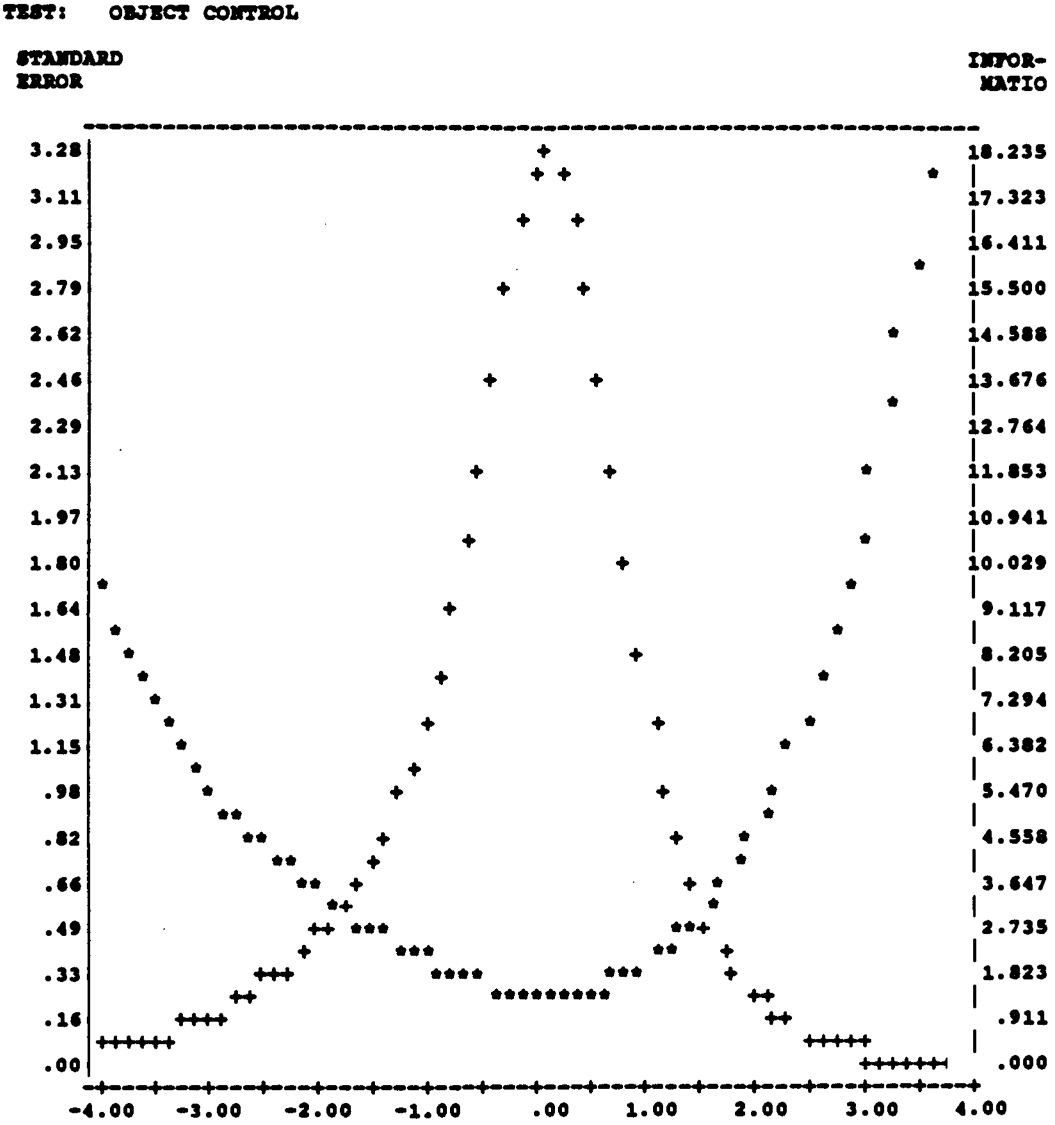


```
MAXIMUM INFORMATION (I) APPROXIMATELY 15.50 AT ABILITY = -1.8571

FOR A NORMAL POPULATION WITH MEAN = .000 AND S.D. = 1.000,
AVERAGE INFORMATION = 10.09 AND RELIABILITY INDEX = .910
```

Appendix Figure 14.

Object Control Subtest Information and Standard Error Curve



```
MAXIMUM INFORMATION (I) APPROXIMATELY 18.24 AT ABILITY = -1.6429
FOR A NORMAL POPULATION WITH MEAN = .000 AND S.D. = 1.000,
AVERAGE INFORMATION = 10.81 AND RELIABILITY INDEX = .915
```

APPENDIX D

Computer Program for Calculating True Scores from Ability Estimates

```
10    REM Emily Cole
20    REM program for thesis, written summer 1989
30    REM This program calculates item response theory true test  scores from given
      ability scores
40    DIM ALOCO(26), BLOCO(26), AOBJ(19), BOBJ(19)
50    FOR I=1 TO 26: READ ALOCO(I)
60    READ BLOCO(I): NEXT I
70       INPUT "Would you like the item parameters printed";P$
100   INPUT"Enter Ability Value  ";THETA
110   FOR I=1 TO 26
120   LL= ((THETA-BLOCO(I))*ALOCO(I))*-1
130   EXL=EXP(LL)
140   PROBL=1/(EXL+1)
150   SUML=SUML+PROBL
160   NEXT I
170   FOR I=1 TO 19: READ AOBJ(I): READ BOBJ(I)
180   LO=((THETA-BOBJ(I))*AOBJ(I))*-1
190   EXO=EXP(LO)
200   PROBO=1/(EXO+1)
210   SUMO=SUMO+PROBO
220   NEXT I
230   TEST=SUMO + SUML
300   LPRINT "Locomotor probability is  ";SUML; "  AT  ";THETA
310   LPRINT "Object control probability is  ";SUMO; "  AT  ";THETA
320   LPRINT " The test probability is  ";TEST; "  AT  ";THETA
322   IF P$="Y" OR P$="YES" THEN 325 ELSE GOTO 600
325   LPRINT "LOCOMOTOR SUBTEST PARAMETERS =  "
330   FOR I = 1 TO 26: LPRINT "A=Parameter =  ";ALOCO(I),"B-     Parameter =
      ";BLOCO(I): NEXT I
340   LPRINT "OBJECT CONTROL PARAMETERS =  "
350   FOR I = 1 TO 19: LPRINT "A-Parameters =  ";AOBJ(I),"B-Parameters =  ";BOBJ(I)
360   NEXT I
450   REM The following are the parameters from the two-parameter Item response
      theory analysis of the TGMD
460   REM parameters were inputted as a-parameter,b-parameter...
500   DATA .946,-6.230,1.469,-1.548,.861,-2.695,1.083,-1.198,1.152, -1.556, .878,-
      3.981,1.984,.272,.761,-2.058,2.564,-.989,2.919, -.082,1.889, .112,1.498, -
      1.291,2.295,-2.46,1.921,-.881,2.415, .633,1.645,-813,1.654,.743,.578, 5.710,
      1.580,-.629,2.522
510   DATA -.613,2.677,-.341,2.963,-.040,.998,-.898,.982,-1.846,.792,2.348,1.066,-401
520   DATA 1.035,-2.097,1.153,-.882,2.424,.305,1.746,.335,2.300,.053,1.583,.128,
      2.261,-.191,.850,-1.828, 1.896,-1.061, 2.249, .036, 3.548,.068,1.637,-1.943,2.836,
530   DATA .363,2.476,.312,1.642,1.133,2.170,-.555,2.754,.147, 1.147,1.739,-
      0.416,2.268,.517
```

APPENDIX E

Appendix Figure 15.

Contingency Tables for Evaluating Mastery

IRT

Locomotor 70% Mastery

		Master	Nonmaster
CTT	Master	446 (.506)	55 (.062)
	Nonmaster	3 (.003)	377 (.428)

(a)

IRT

Locomotor 85% Mastery

		Master	Nonmaster
CTT	Master	267 (.303)	60 (.068)
	Nonmaster	6 (.007)	548 (.622)

(b)

IRT

Object Control 70% Mastery

		Master	Nonmaster
CTT	Master	386 (.440)	8 (.009)
	Nonmaster	23 (.026)	464 (.527)

(c)

IRT

Object Control 85% Mastery

		Master	Nonmaster
CTT	Master	197 (.223	59 (.067)
	Nonmaster	2 (.002)	623 (.707)

(d)

Contingency Coefficients (C) :

Locomotor 70% .93	Locomotor 85% .99
Object Contol 70% .99	Object Control 85% .997

Kappa (k) statistics:

Locomotor 70% .86	Locomotor 85% .98
Object Contol 70% .94	Object Control 85% .99