Trace lines for classification decisions.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road. Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 9432860

Trace lines for classification decisions

Schwarz, Richard Derek, Ph.D.

The University of Arizona, 1994

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

TRACE LINES FOR CLASSIFICATION DECISIONS

by

Richard Derek Schwarz

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

In Partial Fulfillment of the requirement
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

1994

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by Richard Derek Schwarz

entitled Trace Lines for Classification Decisions

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

John R. Bergan — 2-21-94
Date

2-21-94
Date

Darrell Sabers — 2-21-94
Date

2-21-94
Date

Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

John R. Bergan — 2-21-94
Dissertation Director — Date

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under the rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledge of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgement the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: ______________________

Acknowledgements

The roots of education are bitter, but the fruit is sweet.

Aristotle Diogenes Laertius

Many thanks to:

John Bergan - Whose doorstep I arrived at one day, intellectually unswaddled and to whom I owe everything. How many many paths have we travelled together over the years?

Darrell Sabers - Who was always ready to listen (one cannot under estimate the importance of this) and always provided me with the proper advice.

Shitala Mishra - Who influenced my thinking in certain subtle ways and was a benevolent force on my behalf.

Jim Shockey - Who showed me how other social scientists view the world and is one of the nicest people I have ever met.

Dave Thissen - To whom I owe a special thanks for going beyond the call of duty and for providing many timely suggestions regarding this dissertation.

Jessie Fryer - Who supplied me with moral support on occasion when I was needy and for waving her magic wand over a mountain of paperwork.

Dedication

I wish to dedicate this dissertation to my wife, Rose, who has been long suffering not only through this dissertation but also throughout my lengthy career as graduate student.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF ILLUSTRATIONS

FIGURE

ABSTRACT

Referral, placement and retention decisions were analyzed using item response theory in order to investigate several previously unaddressed questions. One question was whether classification decisions could be placed on the latent continuum of ability normally associated with conventional test items. A second set of questions pertained to the existence of differential item functioning (DIF) and testlet functioning (DTF) for the various classification decisions using ethnicity and gender as the grouping variables. Since referral and placement are dependent, two different types of testlets were formed; a referral and placement testlet and a referral, placement, and retention testlet. Test data and educational classification decisions were analyzed for 352 kindergarten children. The resulting "item" parameters were similar to those that might be expected from conventional test items. The *a* parameters where high and positive for both the individual classification decisions and for the testlets indicating adequate discrimination for the various decisions as a function of ability and that the decisions are related to a single underlying variable. The location parameters for the three decisions were low on the ability continuum. The location parameter for placement was lower than the estimate for referral while the estimate for retention was close to the value obtained for placement. Both testlets were graded and had correspondingly low location parameters. Item information was found to be high in the ability range where decisions are made for both individual decisions and testlets. No DIF was found for the Rasch models but was detected for referral for different ethnic groups using the two-parameter model. Using the

Rasch model ignored an important source of DIF contained in the discrimination parameter. DTF was found for the referral and placement testlet when ethnicity was analyzed. Referral decisions for ethnicity were found to be functioning differently for Caucasians versus non-Caucasians. Teachers, in this sample, did not take ability into account when making referral decisions for this group. No DIF was found for placement indicating that evaluation teams did incorporate ability into the decision. Item response theory represents a powerful methodology that could be applied to a variety of new problem types.

Chapter 1

INTRODUCTION

A basic problem often associated with decision-making is how to classify individuals. The problem of classification arises when a decision-maker assesses some individuals and subsequently wishes to assign these individuals to one or more categories based on some observations. The investigator cannot match the individual with a category directly but must infer a category from these measurements. If the classification decisions are correct, then respondents are assigned to the class which they are most likely to belong. It can be assumed that there are a finite number of classes or populations from which the respondent may have come. A respondent can be considered to be a random observation from this population or class. The success of any classification system can judged on the extent to which misclassification is minimized. The traditional approach to examining the classification problem has been to assess how Type I and Type II errors can be minimized and what proportion of the population make up each of the classes. While these are worthwhile questions, they will not be the focus of this dissertation. By using item response theory (IRT), several previously unaddressed questions regarding decision-making can be investigated. One question is to examine what "item" parameters result from analyzing classification decisions as if they were test items using IRT methodology. This allows the classification decision to be empirically placed on the latent continuum and states the decision in probabilistic terms as a function of ability. Secondly, the hypothesis of differential "item" functioning (DIF) for classification decisions can be investigated using IRT methodology. DIF in this context evaluates the validity of classification decisions after they have occurred. The type of classification problems addressed in this dissertation are special education referral and placement and retention in grade.

Classification decisions can be investigated in the same manner normally associated with test items or responses to rating scales using latent structure models such as IRT. Thissen (1993) suggests that researchers can redefine what constitutes an item in order to pursue new questions. Since classification decisions are categorical and the underlying dimension can be conceptualized as ability, it seems appropriate that they be examined in the manner usually associated with test items. The probability of a positive response (i.e., getting the item correct) or not being referred, placed, or retained are all a function of ability. Children who are high in basic skills (i.e., math and reading ability) are not likely to be referred or placed in special education or be retained (Bergan, Sladeczek, Schwarz, & Smith, 1989). Classification decisions can be modeled using the trace line of the two-parameter logistic model (Birnbaum, 1968). A trace line shows how the probability of a decision varies as a function of ability. The "item" statistics (i.e., the location and discrimination parameters) allow the classification decision to be reported on the same scale as ability. The *a* parameter, which is the slope of the trace line, tells whether the classification decision discriminates well between two groups. A high positive value for the *a* parameter indicates that a decision discriminates well between placed and nonplaced groups, for instance. The location parameter (i.e., b parameter) is the point on the ability scale where a child has 50% chance of not being either referred, placed or retained. For classification decisions, the location parameter is expected to be low on the ability continuum. The location parameter for placement is hypothesized to be lower on the ability continuum than referral since it is assumed that children who were placed are lower on the ability continuum then children who were just referred.

An important factor affecting the discriminating power of a test is the scatter of the responses around a particular level of the latent variable denoted by θ. IRT provides an estimate of the dispersion of the total likelihood that can be used to evaluate the precision of the maximum likelihood estimates around a given level of θ. This measure of information at each interval on the latent continuum is called item information. Item information can be used to evaluate the effectiveness of a classification decision. Decisions are more adequately characterized by item information than reliability. Wainer (1985) suggests that the usual measures of reliability such as coefficient α are often inappropriate since they characterize the entire test, rather than the area of the latent continuum where decisions are made. Item information can be used to ensure that information is maximized in the region of the decision. Since decisions regarding special education referral and placement and retention in grade occur low on the ability continuum, information should be maximal in this region. If item information is low in the region of the decision, this would indicate that classification is error prone.

From a validity standpoint, one of the primary concerns with classification decisions has been the perception that ethnic minorities are over represented in special education classes (Reschly, 1988; Shepard, 1989). The cause for this over representation is alleged to be "biased" tests, especially IQ tests. Angoff (1993) suggests that researchers apply DIF technology to new contexts. Classification decisions represent one such new context. The hypothesis of differential classification decisions for various subpopulations can be investigated using differential item functioning methodology (Lord, 1980; Thissen, Steinberg, & Wainer, 1993; Kelderman & Macready, 1990). DIF refers to an *unexpected* difference in item performance between two groups of examinees that have been matched on the trait being measured by the test (Dorans & Holland, 1993).

A DIF analysis begins with the specification of a reference group and a focal group. The reference group is the standard against which the focal group is compared (Holland & Thayer, 1988). The "studied item" is the focus of the DIF analysis. A DIF analysis using IRT can be simplified with the selection of a designated anchor that is iteratively purified (Thissen, Steinberg, & Wainer, 1993). This process consists of eliminating items from the anchor that demonstrate DIF until no DIF exists. Thissen et al. found that anchors consisting of one to four items work well. The designated anchor can then be used to test all the other items for DIF.

The conditional probability of a correct response given a specified proficiency level of a respondent is given by the item trace line. The extent to which the item trace lines differ between the focal and reference groups can be construed as demonstrating DIF (Lord, 1980; Thissen, Steinberg, & Wainer, 1993). In statistical terms, the null hypothesis is one of no DIF. Using Thissen and Steinberg's (1988) methodology, statistical fit is evaluated for the restricted model obtained by imposing equality constraints on the parameters for the studied item across both the reference and focal groups. This constrained model may be compared with an unconstrained model using the likelihood ratio statistic. The likelihood ratio statistic for the unconstrained model can be subtracted from the constrained model, which results in difference a chi-square. If the resulting difference chi-square is not significant, then the null hypothesis of no DIF is not rejected. In this dissertation, the hypothesis of DIF will be investigated for gender and ethnicity. These variables were chosen because they are widely recognized as critical variables in the study of DIF. DIF with respect to gender and ethnicity will be examined for referral and placement, and for retention.

Referral and placement is a two-step process. First, a child is referred and then a decision is made as to placement. Since referral and placement is a two-step process, it might be advantageous to model it as a two-step procedure. Secondly, referral and placement are highly dependent because a child must first be referred before he or she is placed. If a child is placed, then it is extremely likely that the child had already been referred. In addition, if a child is referred it is very likely that he or she will be placed. This violates the assumption of local "item" independence. Local independence suggests that the chance of "success" on one "item" should not be affected by the chance of "success" on another item. Fortunately, DIF technology can be extended to a more macro unit of test analysis known as testlets (Wainer & Kiley, 1987; Rosenbaum, 1988; Thissen, Steinberg, & Mooney 1989; Wainer, Sireci, & Thissen, 1991) that allows the assumption of local independence to be met. Thissen, Steinberg, and Mooney (1989) used the summed score of the items within the testlet as a sufficient statistic for scoring the test. Bock's (1972) nominal model was used to fit the testlets. Thissen et al. (1989) were able to show local independence among the testlets with little loss of information compared with each test item being scored independently. Since referral and placement decisions are not independent, it is more appropriate to analyze the referral and placement process using testlets. The categories comprising the testlet will be the summed score of no referral or placement, just referral, and placement and referral. Since these can be thought as being ordered, Samejima's (1969) graded response or Master's (1982) partial credit model could be used to fit the testlet. Referral, placement, and retention decisions also form a testlet. The four categories comprising this testlet are the summed score of: no referral, placement or retention; just referral or retention; referral and placement or referral and retention; referral, placement, and retention. Differential testlet functioning is tested and interpreted in the same manner as differential item functioning. The results from the item analysis can be compared to the testlet analysis to determine the degree of agreement between the two procedures.

Chapter 2

THEORETICAL BACKGROUND

This chapter presents a brief overview of latent structure models in which latent class models are compared with models that assume a continuous latent variable such as IRT. It is hoped that such a comparison will facilitate understanding of latent structure modeling. The main body of this chapter includes a discussion of various IRT models and a section on estimation. The second section presents an overview of differential item functioning.

Latent Structure Models

Latent structure analysis is a method for analyzing multivariate categorical or continuous data when one or more unobserved variables is said to account for the observed relationships among the variables. One objective of latent structure analysis is to give the most parsimonious explanation for the interrelationships that exist among the observed data. Latent structure analysis can be viewed as a data reduction method through model fitting. The number of latent variables T is usually much less than the number of manifest variables. Another objective is to locate individuals in the latent space on the basis of their observed score. A particular latent structure model is a specification of the nature of the latent space. The probability of positive response to any item is a function of the respondents' position in this latent space. For every individual in the sample, there is an associated value of the latent variable. The variation in the latent variable reflects individual differences. Frequently, latent variables are conceptualized as being continuous. However, sometimes the latent variable can be thought of as being discrete and comprised of a finite number of classes. All members of the sample fall into one of the mutually exclusive classes and have the same value of the latent variable. The assumption basic to both these models and all other latent variable models is the axiom of local independence.

Local Independence

In IRT, the observed variables, Y_j's, are discrete but the latent variable, X, is continuous while in latent class models both the manifest and latent variable are discrete. These two methods appear to be unrelated but both have a common reliance on the axiom of local independence or conditional independence. Local independence implies that the observed relationships among the manifest variables is accounted for by their common dependence on one or more latent variables. Local independence states

$$P(x_{i1},...,x_{ik}|\theta) = P(x_{i1}|\theta)\cdots P(x_{ik}|\theta) \tag{1}$$

where θ represents the latent variable and x_{ik} denotes the response on item i for respondent k. If the data are Bernoulli random variables then $f(x_{ik}|\theta)$ can be written as

$$f(x_{ik}|\theta) = \begin{cases} p_k(\theta) & for\ x_{ik} = 1 \\ 1 - p_k(\theta) & for\ x_{ik} = 0. \end{cases}$$

Then Equation (1) can be written as a two-component multinomial mixture

$$f(x_{i1}, ..., x_{ik}|\theta) = p_1(\theta)^x_{ik}[1 - p_1(\theta)]^{1-x_{ik}} \cdots p_k(\theta)^{x_{ik}}[1 - p_k(\theta)]^{1-x_{ik}}. \qquad (2)$$

Consider the case where θ takes only t distinct values $\theta_1, ..., \theta_t$. Let φ_v be the probability that a randomly sampled respondent has value θ_v of the latent variable. Using Equation (2) the marginal distribution of the random variables $X_{i1}, ..., X_{ik}$ is

$$f(x_{i1}, ..., x_{ik}) = \sum_{v=1}^{T} p_1(\theta_v)^{x_{i1}}[1 - p_1(\theta_v)^{1-x_{i1}}] \cdots p_{kv}[(\theta)^{x_{ik}}(1 - p_k(\theta_v)^{1-x_{ik}}]\varphi_v. \quad (3)$$

This is equivalent to a latent class model (Andersen, 1988). Suppose that a cross-classification is observed with respect to four qualitative variables; A, B, C, and D. The basic assumption of latent class modeling is the conditional or local independence of variables A,B,C, and D given the existence of a latent variable θ made of T latent classes. A significant χ^2 would indicate a association among the four variables. If two latent classes are hypothesized to exist and the likelihood statistic fits the data then the latent types explains the association among the variables.

If the distribution of θ is continuous and described by the density function $\varphi(\theta)$ then the marginal distribution is given by

$$f(x_{i1}, ..., x_{ik}) = \int f(x_{i1}|\theta) \cdots f(x_{ik}|\theta)\varphi(\theta)d(\theta). \qquad (4)$$

Various models which have been proposed in the literature that arise when different latent densities are chosen for $\varphi(\theta)$ and for the conditional probability functions in (4). Lord (1952), Bock and Liberman (1970), Bock (1972), Christofferson (1975) followed by Muthen (1978), use the normal density. Birnbaum in Lord and Novick (1968) treated the unidimensional case using the logit for conditional probabilities and the normal density for the latent variable. Ability is frequently considered to be normal and serves as the underlying latent dimension for many IRT models. Bartholomew (1980) proposed the use of the logit for both functions.

Item Response Theory

IRT is an efficient method of ordering examinees along a latent continuum such as ability. The origins of item response theory can be traced back to Richardson (1936) and Lawley (1943). Richardson provided a method for obtaining item response parameter estimates while Lawley (1943) defined new procedures for parameter estimation. Lord (1952), considered by many to be the father of IRT, proposed the two-parameter normal ogive model in his doctoral dissertation. Rasch (1960) developed his own IRT model independently which bears his name. Birnbaum (1968) substituted the logistic form for the normal ogive which made parameter estimation more tractable. Bock and Aitkin (1981) solved many of the parameter estimation problems using an alternative derivation of the EM-algorithm called marginal maximum likelihood.

In order to fully specify the probabilities for an IRT model, a suitable latent density $\varphi(\theta)$ must be chosen. In the case, where θ is a continuous variable, the probabilities $\theta_1, ..., \theta_t$ for each of the latent classes are replaced by a latent density $\varphi(\theta)$ which can be normal (Lord and Novick, 1968; Bock, 1972). This latent density describes the variation of θ over the given population with mean μ and variance σ^2. Bartholomew (1980) discussed other choices for the latent density including the inverse Cauchy distribution and gave criteria for how to choose $\varphi(\theta)$. He argues that the logistic density

$$\varphi(\theta) = \frac{1}{\sigma} exp(\frac{(\theta - \mu)}{\sigma}/[1 + exp\frac{(\theta - \mu)}{\sigma}]^2 \tag{5}$$

should be preferred based mostly on ease of numerical computation. Muthen (1978), (1979), Christofferson (1975) Muthen and Christofferson (1981) gave models, where θ is a multivariate latent variable, $\varphi(\theta)$ the multivariate normal density and the responses probabilities $\pi_i^S(\theta)$ are generalized probit functions. In many applications of IRT, it is assumed that only one latent variable is needed to account for the interdependcies among responses. These models are referred to as unidimensional models. Multidimensional models (Reckase, 1985) have recently received much greater attention.

Let $p_{ijkl}(\theta)$ be a response vector indicating the probability of observing an individual in cell *(ijkl)* given a value of the latent variable θ. For the four dichotomous variables (i.e., items) A,B,C,D denoted as

$$p_{ijkl}(\theta) = \pi^{\overline{A_i}\theta}\pi^{\overline{B_j}\theta}\pi^{\overline{C_k}\theta}\pi^{\overline{D_l}\theta} \tag{6}$$

where $\pi^{\overline{A_i}\theta}$ is the probability that variable A is at level i given the value of the latent variable θ, and $\pi^{\overline{B_j}\theta}$ is the probability that variable B is at level j given the value of latent variable θ. Similar definitions can be given for the variables C and D. For a variable A, for example, there are two conditional probabilities $\pi^{\overline{A_1\theta}}$ and $\pi^{\overline{A_2\theta}}$ for correct and incorrect responses but only one needs to be specified since $\pi^{\overline{A_2\theta}} = 1 - \pi^{\overline{A_1\theta}}$. The response probabilities $\pi^{\overline{A_i\theta}}, ..., \pi^{\overline{D_l\theta}}$ are functions of θ known as item characteristic curves. The respondent's probability of a positive response is assumed to be a monotonically increasing function of θ. Item characteristic curves can be used to model referral, placement, and retention decisions.

Integrating over the latent density, the marginal cell probabilities p_{ijkl} are obtained as

$$p_{ijkl} = \int \pi^{\overline{A_i\theta}} \cdot \ldots \cdot \pi^{\overline{D_l\theta}}. \qquad (7)$$

Model construction consists of specifying a functional form for the conditional probabilities and specifying a latent density.

Some IRT models

There are several functions commonly used for modeling the response probabilities. Most models use the logistic function, g(x) = ln(x/1-x), and the probit function, $g(x) = \Phi^{-1}(x)$. The simplest model using a logistic function is the well-known Rasch (1960) or one-parameter model for item i

$$\pi^{A_2\theta} = \frac{1}{1+exp(\theta-b_i)} \qquad (8)$$

where b_i refers to the location or difficulty parameter for item A. It represents the point on the latent scale where the respondent has a 50 percent probability of a positive response on item A. The response probabilities for the Rasch model for the two levels of responding can be written as

$$\begin{aligned}
\pi^{A_1}(\theta) &= & \frac{exp(\theta - b_i)}{1 + exp(\theta - b_i)} \\
\pi^{A_2}(\theta) &= 1 - \pi^{A_1}(\theta) = & 1/1 + exp(\theta - b_i) \\
&\vdots & \\
\pi^{D_1}(\theta) &= & \frac{exp(\theta - b_l)}{1 + exp(\theta - b_l)} \\
\pi^{D_2}(\theta) &= 1 - \pi^{D_1}(\theta) = & 1/1 + exp(\theta - b_l).
\end{aligned}$$

For a Rasch model the probability of response vector *ijkl* for a respondent is

$$p_{ijkl}(\theta) = exp(b_i z(i) + ... + b_l z(l) + t\theta)/[(1 + e^{(\theta - b_i)}...(1 + e^{(\theta - b_l)}] \tag{9}$$

in which z(1) = 1, z(2) = 0 and t=z(i)+...+z(l). The score, t=z(i)+...+z(l), is a sufficient statistic for θ. The number correct for an item is sufficient statistic for the item difficulty. The existence of sufficient statistics is one advantage of the Rasch model. Other advantages include the estimation of fewer parameters than other IRT models and the attainment of specific objectivity (Rasch, 1966). This property permits the separate estimation of the location and ability parameters. The Rasch model could be used to estimate where a classification decision is located on the latent continuum of ability.

There have been a number of extensions of the Rasch model to handle polytomous data (Andrich, 1978; Masters, 1982). Andrich modified the Rasch model in order to analyze rating scales. Masters (1982) developed a another variation on the Rasch model that he referred to as the partial credit model. Let w,x,y,z to be m = *0,..t* ordered response categories to an item where the elementary relations 0 < 1; 1 < 2; 2 < 3 completely capture the intended relationships. The probabilities $\pi^{\overline{w_i}\theta}, ..., \pi^{\overline{z_l}\theta}$ represent a response category at the *ijkl* level for a given value of θ. It is assumed that as θ increases that a response in the i category will decrease. The probability of a positive response is governed two parameters: the level of the latent variable and the transition parameter δ_m associated with the step between adjacent categories for a particular item. The transition parameters correspond to the intersections of adjacent probability curves. Only *T-1* transitions are required to uniquely define the order. The probability of responding 1 rather than 0 on alternative w is

$$\pi^{\overline{w_i}\theta} = \frac{exp(\theta - \delta_1)}{1 + exp(\theta - \delta_1).} \tag{10}$$

This procedure can be applied to each pair of adjacent response categories *s-1* and s. The odds of responding in category 1 rather than in category 0 on an item or observation s

$$\pi^{\overline{x_j}\theta} / \pi^{\overline{w_i}\theta} = exp(\theta - \delta_1). \tag{11}$$

The odds of responding in category 1 rather than in category 2 on an item or observation s is

$$\pi^{\overline{y_k}\theta}/\pi^{\overline{x_i}\theta} = exp(\theta - \delta_2). \tag{12}$$

It follows that the odds can be written in this way for any pair of response categories. For example, the odds of responding in category y rather then category w for an item s is

$$\pi^{\overline{y_k}\theta}/\pi^{\overline{w_i}\theta} = exp(2\theta - \delta_1 - \delta_2). \tag{13}$$

With the usual constraint that response categories sum to one

$$\begin{aligned} \pi^{\overline{w_2}\theta} &= 1/\Phi \\ \pi^{\overline{x_1}\theta} &= exp(\theta - \delta_1)/\Phi \\ \pi^{\overline{y_1}\theta} &= exp(2\theta - \delta_1 - \delta_2)/\Phi \\ \pi^{\overline{z_1}\theta} &= exp(3\theta - \delta_1 - \delta_2 - \delta_3)/\Phi \end{aligned} \tag{14}$$

where Φ is the sum of the numerators to ensure that the response probabilities sum to one. These four expressions can be captured in a single expression which gives the probability of a person scoring in category t rather than category *t-1* on an item as a function of the parameter θ and item parameter δ_m

$$\pi^{\overline{w_i}\theta}, ..., \pi^{\overline{z_l}\theta} = \frac{exp\sum_{r=0}^{x}(\theta-\delta_m)}{\sum_{s=0}^{t-1} exp\sum_{r=0}^{s}(\theta-\delta_m)} \quad x = 0, ...t - 1. \tag{15}$$

Thus for each item there are a series of t logistic curves that differ in terms of their transition parameters. The partial credit model could be used to model the classification testlets if it was important to maintain specific objectivity. Master (1988) suggests that an additional advantage to the partial credit model is that the transition parameters are defined locally allowing for a more convenient interpretation when it comes to studying the operational definition of variables.

In the Rasch model, the slope parameters for the items were equal. The Rasch model can be extended by allowing the response probabilities to depend on a third parameter a called item discrimination which is proportional to the slope when $\theta = b$. The a parameter reflects the rate of change in a positive response as a function of θ. The resulting two-parameter logistic model (Birnbaum, 1968) for item s is

$$\pi^{\overline{S_1}\theta} = exp(\theta - b_s)a_s/[1 + exp(\theta - b_s)a_s]. \tag{16}$$

The two-parameter model could be used to model classification decisions to determine how the location and discrimination parameters vary as a function of ability.

Lord (1952) proposed a two-parameter model known as the normal ogive

$$\pi^{\overline{S_1}\theta} = \Phi((\theta - b_s)a_s) \tag{17}$$

in which Φ is the cumulative distribution function for the normal standard distribution. A scale factor of 1.7 can be added to the logistic form to achieve the normal metric. Guessing by an examinee can be modeled by the lower asymptote which represents the probability that low ability examinees will answer the item correctly. It does not appear in the logit but forms an additive and multiplicative factor. The addition of guessing to the two-parameter model results in the three parameter model.

Samejima (1969) used the two-parameter models to handle the case where item responses are contained in two or more ordered response categories. A two-parameter logistic is produced for each possible response alternative to represent the probability of an individual selecting that category or higher. The operating characteristic of the graded response model can be defined as the probability of a response for a given level of θ in or above category y for categories w,x,y,z, as

$$\frac{1}{1+exp[1+-a(\theta-b_{y-1})]} - \frac{1}{1+exp[1+-a(\theta-b_y)]} \tag{18}$$

$$= P_{(y-1)} - P_{(y)}. \tag{19}$$

This model is sometimes referred to as a difference model since the probability of choosing the y^{th} ordered category is the difference in the probabilities of the category characteristics associated with the y^{th} and $y^{th}-1$ categories. The b parameter is defined as the point on the latent continuum where the probability is 50% that a response is in category y or higher. The a parameters are generally assumed to be equal for responses $m = 0,...t$. The shape of the item characteristic curve ICC will generally be non-monotonic except when $m = 0$ or when $m = t$.

Estimation in IRT

There are many methods of estimating the parameters in IRT such as conditional maximum likelihood, joint maximum likelihood, and marginal maximum likelihood (MML). The estimates arrived at by conditional maximum likelihood are not consistent (Andersen, 1972) since the bias fails to disappear when the sample size increases.

One solution to this problem is to condition the ability parameters out of the likelihood equation by replacing them with their minimally sufficient statistics. The likelihood equation is then estimated conditional upon the sufficient statistics. Since sufficient statistics exist for the Rasch model, this conditional approach can be used. Consistent estimates will not be obtained for other models because sufficient statistics do not exist. Joint maximum likelihood suffers from estimates of ability that can be biased which can in turn lead to biased estimates of the item parameters. Secondly, for models without sufficient statistics, the number of nuisance (ability) parameters that have to be estimated increases with sample size. These nuisance parameters cannot be eliminated by conditioning them out. One proposed solution was to specify a fixed number of ability points. This is difficult to justify on the grounds that item estimates may become biased when the numbers of items is small and the failure of a finite number of points in representing the distribution of ability in the population. Lord (1986) suggests that marginal maximum likelihood is an improvement over joint maximum likelihood because item parameters and ability parameters do not have to simultaneously estimated. Bock and Liberman (1970) made a fundamental advance by integrating over the parameter (latent) distribution and estimating the structural parameters (items) in the marginal distribution which became known as marginal maximum likelihood (MML). Bock and Aitkin (1981) improved on Bock and Liberman by applying an alternative derivation of the EM-algorithm. The EM-algorithm of Demspter, Laird, and Rubin (1977) is a two-step process consisting of a expectation step (E step) and maximization step (M step) that allows the computation of maximum likelihood estimates from incomplete data. The EM-algorithm is an iterative procedure ideally suited to estimation in latent structure models. During the E-phase of each iteration the incomplete data are completed by estimating them on the basis of the observed data and the provisory parameter estimates obtained in the previous iteration. During the M-phase of each iteration the unknown parameters are estimated again on the basis of the complete data arrived at in the E-phase. These last two steps are repeated until the maximum likelihood estimates converge to some specified criterion or the maximum number of iteration steps allowed is exceeded. Thissen (1982) demonstrated the use of MML for the Rasch model.

Bock and Liberman (1970) took a very different approach to the problem of incidental parameters than previous estimation procedures. Bock and Liberman proposed an alternative procedure in which ability estimates are removed in order to estimate item parameters unconditionally by integrating them out of the likelihood function. The basic element in this approach is to estimate the structural (i.e., item) parameters by ML in the marginal distribution obtained by integrating over the distribution of the incidental (ability) parameters. The resulting likelihood function is referred to as the marginal likelihood function. Because there are 2^n possible patterns for a vector of item responses, the Bock and Liberman approach is practical only for a few items, say n < 12. Despite this limitation, the initial conceptualization allowed a reformulation to be made by Bock and Aitkin (1981) that made computation feasible for a wide variety of problems.

The Bock and Aitkin Approach

Bock and Aitkin (1981) reformulated the Bock and Liberman approach in a manner that can be derived as extension of the missing information principle of the EM algorithm (Dempster, Laird, & Rubin 1977). Since no sufficient statistics exist for θ, each individual observation θ_i is replaced by its conditional expectation given the observed response vector, x_i, for a respondent i. The conditional distribution of θ given that $x = x_i$ from Bayes' theorem is

$$g(\theta|x_i) = \frac{P(x=x_i)|\theta)g(\theta)}{P(x=x_i)} \quad . \tag{20}$$

Using the local independence assumption for item j, the probability of a response pattern x_i conditional on θ

$$P(x = x_i|\theta) = \sum_{j}^{n}[\Phi_j(\theta_i)]^{x_{ij}}[1 - \Phi_j(\theta_i)]^{1-x_{ij}}. \tag{21}$$

The unconditional probability for a randomly selected individual from a population with a continuous ability distribution $g(\theta)$ is

$$P(x = x_i) = \int_{-\infty}^{\infty} P(x = x_i|\theta)g(\theta)d\theta. \tag{22}$$

The conditional expectation of θ given $x = x_i$ using relations (21) and (22) is

$$E(\theta|x_i) = \frac{\int_{-\infty}^{\infty} \theta g(\theta) \prod_j^n [\Phi_j(\theta)]^{x_{ij}} [1-\Phi_j(\theta)]^{1-x_{ij}} \, d\theta}{\int_{-\infty}^{\infty} g(\theta) \prod_j^n [\Phi_j(\theta)]^{x_{ij}} [1-\Phi_j(\theta)]^{1-x_{ij}} \, d\theta} \quad . \tag{23}$$

Since intergrals are difficult to evaluate, methods such as numerical quadrature have been developed for approximating them to any desired degree of accuracy. A continuous distribution with finite number of moments can be approximated by a discrete distribution, such as a histogram, over a finite number of points. Using a quadrature approximation, the problem of finding the sum of the area underneath a density function is simplified by the problem of finding the sum of the areas of a finite number of rectangles which approximate the area under the curve. The midpoint of each histogram on the ability scale, $X_k (k = 1, 2, ..., q)$, is called a node. Each node has an associated weight $A(X_k)$ which indicates the height of the density function in the neighborhood of X_k and the width of the histogram. Tables of X_k and the corresponding weights $A(X_k)$ for approximating the Gaussian distribution are given by Stroud & Sechrest (1966). Using Gauss-Hermite quadrature the probability in Equation (22) can be approximated by the sum

$$\sum_k^q P(x = x_i|X_k) A(X_k). \tag{24}$$

Equation (23) can now be rewritten approximating the intergrals using q points indexed by k with the *i-th* respondent to the *l-th* score pattern

$$E(\theta|x_l) \approx \frac{\sum_k^q X_k L_l(X_k) A(X_k)}{\tilde{P}_l} \tag{25}$$

where $\tilde{P}_l = \sum_k^q L_l(X_k)A(X_k)$ and $L_l(X_k)$ is the conditional probability of X_l given that $\theta = X_k$.

Equation (25) results in a weighted mean of the X_k. Since there are s score patterns there are s values of $E(\theta|x_l)$. The number of responses at x_l is r_l. For the *j-th* item, the number correct is x_{lj}. A probit model is then fitted to s points using $E(\theta|x_l)$ as the expected ability variable, with $x_{lj}r_l$ as the number correct out of r_l for this level of ability. The expected frequency of correct responses to the *j-th* item and at a given level of θ is

$$\overline{r_{jk}} = \frac{\sum_l^s r_l x_{lj} L_l(X_k)A(X_k)}{\tilde{P}_l} \tag{26}$$

and the expected number of respondents at that level is

$$\overline{N}_k = \frac{\sum_l^s r_l L_l(X_k)A(X_k)}{\tilde{P}_l} \tag{27}$$

in which the sum of $\overline{N_k}$ is *N.*

The EM algorithm is a numerical procedure that consists of two steps per item per cycle. The process begins with the expectation (E)-step, in which the provisional estimates of the item parameters are obtained by computing $L_l(X_k), k = 1, 2, ...q$, and $\tilde{P}$ for response patterns *l, (l = 1,2,...,s).* These two values are accumulated to compute for each item the "expected" number of examinees $\overline{N_k}$ at each quadrature point and the number of these $\overline{r_{jk}}$ correctly responding to an item. The second stage is the maximization (M)-step, in which the improved estimates of the item parameters are obtained by performing a probit MLE on $\overline{N_k}$ and $\overline{r_{jk}}$ where X_k is the independent variable weighing the corresponding term by $A(X_k)$. The algorithm cycles until so convergence criteria are met or the maximum number of iterations specified is exceeded. For the Rasch model, the marginal likelihood equation is for the difficulty parameter is

$$\sum_{k}^{q} \overline{r_{jk}} A(X_k) - \sum_{k}^{q} \overline{N_{jk}} \Phi_j(X_k) A(X_k) = 0 \tag{28}$$

where Φ is the one parameter logistic.

The likelihood ratio chi-square statistic used to evaluate the fit of a particular model is

$$G^2 = 2 \left(\sum_{t}^{s} r_l \, log_e \, \frac{r_l}{N\tilde{P}_l} \right) \tag{29}$$

with *s* - *2n* degrees of freedom ignoring pattern counts with $r_l = 0$. If 2^n is large relative to N, the pattern counts will be sparse and estimation will become unstable.

Assuming the item parameters have been estimated, they are considered as known and the respondent's θ values can be estimated as a separate process either by maximum a posteriori (MAP) or Bayesian expected a posteriori (EAP) estimation.

Information

An estimate of the dispersion of the total likelihood when normally distributed may be used to specify the precision with which the MLE(θ) estimates θ. As a byproduct of ML estimation, the estimated standard error is equal to the negative inverse of the expected value of the second derivatives of the loglikelihood (Fisher, 1925). Since the precision of the estimates vary with the width of the total likelihood, the standard error of estimation given by

$$SE(\theta_o) = \frac{1}{\sqrt{I(\theta_o)}} \tag{30}$$

which is inversely related to the amount of information provided by a test at a given level of θ. Information in this context is equivalent to Fisherian information (Fisher, 1925). Information by this definition is approximately equal to $1/(S.E.)^2$ and reflects how much is known at about a parameter at that level of θ. Item information for the two-parameter logistic is

$$I(\theta) = \frac{P'(\theta)^2}{P(\theta)[1-P(\theta)]} \tag{31}$$

in which $P'(\theta)^2$ is the first derivative of $P(\theta)$ with respect to θ. Test information is simply the sum of the individual item information functions. It is desirable to have high information in the area of the decision. If information is low in the area of the decision then the decision-making process is probably error prone.

Differential Item Functioning

A psychometric difference in how an item functions for two matched groups is referred to as differential item functioning (DIF) (Dorans & Holland, 1993). Shepard (1982) defined DIF as psychometric differences that misrepresent the competence of one group. DIF then can be viewed as a source of invalidity. Lord (1980) provided the earliest IRT definition of DIF.

> If each item in a test had exactly the same item response function in every group, then people of the same ability or skill would have exactly the same chance of getting the item right, regardless of their group membership. Such a test would be completely unbiased. If on the other hand, an item has a different item response function for one group than for another it is clear the item is biased. (p. 212)

The more neutral term "differential item functioning" is preferred to the older term "item bias" since the older term does not always accurately reflect the circumstances. Bias is a social judgement while DIF is defined statistically. Dorans and Holland (1993) suggest it is important to distinguish between DIF and impact. Impact refers to a difference in performance between two groups that can be explained by stable consistent differences in ability across groups. Dorans and Holland give the examples of Asian-Americans scoring higher on the SAT-mathematics than other groups or high school seniors scoring higher than juniors. DIF, however, refers to an unexpected difference in performance between two groups who have been matched on the trait being measured on the test.

In a DIF analysis, the performance of the focal group is of primary interest when compared to the reference group. The notion that for given a level of ability members of the focal group are compared to the reference group is basic to most DIF procedures. There are many methods with which to test the hypothesis of DIF. Chief among these procedures are chi-squared based methods such as the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), the standardization approach (Dorans & Kulick, 1986), methods based on latent class analysis (Kelderman & Macready, 1990) and IRT (Lord, 1980; Thissen, Steinberg, & Wainer, 1988). What follows is a brief overview of these methods of DIF detection.

Mantel and Haenszel (1959) introduced a measure based on chi-square in order to study matched groups. The Mantel-Haenszel (MH) chi-square tests the null hypothesis of

$$H_o : \frac{p_{R_j}}{q_{R_j}} = \frac{p_{F_j}}{q_{F_j}} \tag{32}$$

in which j = 1,...K, and

p_{R_j} = proportion correct in reference group

q_{R_j} = proportion incorrect in reference group

p_{F_j} = proportion correct in focal group

q_{F_j} = proportion incorrect in focal group

j = the j^{th} matched set.

The alternative hypothesis, known as the constant odds ratio hypothesis, is stated as

$$H_1 : \frac{p_{R_j}}{q_{R_j}} = \alpha \frac{p_{F_j}}{q_{F_j}} \tag{33}$$

where the parameter α is the odds ratio under H_1. The odds ratio in the K 2 × 2 tables is

$$\frac{p_{R_j}}{q_{R_j}} / \frac{p_{F_j}}{q_{F_j}} = \frac{p_{R_j} q_{F_j}}{P_{F_j} q_{R_j}} \tag{34}$$

The Mantel-Haenszel chi-square statistic is based on

$$\frac{\sum_j (A_j - \sum_i E(A_j))^2}{\sum_j Var(A_j)} \tag{35}$$

where

$$Var(A_j) = \frac{n_{R_j} n_{F_j} m_{1j} m_{0j}}{T_j^2 (T_{j-1})} \tag{36}$$

and

A_j = the number of respondents in the reference group who answered correctly

n_{R_j} = number of respondents in reference group

n_{F_j} = number of respondents in focal group

m_{1j} = marginal for proportion correct

m_{0j} = marginal for proportion incorrect.

T_j = total number of reference and focal group members in the j^{th} matched set.

Under H_o, Mantel-Haenszel has an approximate chi-square distribution with a single degree of freedom. When α is equal to one, the the alternative hypothesis reduces to the null hypothesis of no DIF. One problem with the MH statistic is that it only supplies information about the intercepts of the response functions (Bock, 1993). It gives no information that might indicate that the slopes are different. Its main advantage lies in not being iterative and can be computed quickly in order to screen a large number of items for DIF.

The standardization procedure is another method for DIF detection. An item exhibits DIF, according to the standardization procedure, when the expected performance on an item differs for matched groups (Dorans & Holland, 1993). The standardization procedure compares the nonparametric item test regression for the focal and reference groups. Visual analysis is an important component of this approach. Let $E_f(I|M)$ define the item test regression for the focal group and similarly let $E_r(I|M)$ be the item test regression for the reference group where I is the item score variable and M is the matching variable. The definition of DIF using the standardization approach is $E_f(I|M) \neq E_r(I|M)$. An alternative definition of DIF is

$$D_m = E_{fm} - E_{rm} \tag{37}$$

where E_{fm} and E_{rm} are item test regression for the focal and reference groups for a score level m. The D_m are the fundamental measures of DIF according to the standardization approach. A numerical index called the standardized p-difference is used to indicate DIF that can range from -1 to +1. Positive values of this index indicate that the item favors the focal group whereas negative values favor the reference group. Values less than -.05 and +.05 are considered negligible. Bock (1993) suggests that for a test containing a large number items, the standardization procedure and the IRT approach are essentially equivalent since the observed regression of the item score on test score is essentially an item response function. IRT procedures are much more demanding since MML estimation is used and all item regressions are described with the same family of response functions.

Latent-class modeling has also been applied to the question of differential item functioning when the latent variable is conceived as being categorical (Kelderman & Macready, 1990). Kelderman and Macready (1990) considered a variety of models for the detecting DIF and suggested that this class of models is especially appropriate when the latent variable exists in two mutually exclusive and exhaustive states such as master or nonmaster. They assumed that two types of error responses occur; omission errors and intrusion errors. Omission errors occur when masters miss items while intrusion errors occur when nonmasters respond correctly to items. They suggest that an item exhibits DIF in mastery modeling if the omission and intrusion rates differ across groups with respect to the grouping variable. DIF detection in the case of latent class models, like IRT models of DIF, requires the comparison of likelihood statistics between unconstrained models and those imposing parameter constraints.

Two basic methods exist within IRT for assessing DIF. One method devised by Lord (1980) consists of testing the significance of the difference between item parameters for the focal and reference groups that is often referred to as Lord's chi-square. He proposed the test statistic

$$z = (b_R - b_F)/SE(b_R - b_F) \tag{38}$$

where

$$SE(b_R - b_F) = \sqrt{V(b_R) + V(b_F)}. \tag{39}$$

As a byproduct of MML estimation, the second-derivative approximations of the standard errors of the item parameters can be obtained. The test statistic given in Equation (38) can be referred to the standard normal table or squared and referred to the chi-square distribution with one degree of freedom. Lord also proposed a simultaneous test that $a_R = a_F$ and $b_R = b_F$ that is based on the Mahalanobis distance (D^2) between the parameter vectors for two groups (Thissen, Steinberg, & Wainer, 1988). Mahalanobis distance for this case is

$$D^2 = v'\Sigma^{-1}v \tag{40}$$

where Σ is the covariance matrix of parameters and v is a vector of parameters. D^2 is distributed as a chi-square with two degrees of freedom for the null hypothesis. Thissen, Steinberg, and Wainer (1988) suggest that with the advent of MML estimation procedures that allow simultaneous estimation of item parameters for two or more groups, Lord's procedures are no longer really necessary.

Thissen, Steinberg, and Gerrad (1986) proposed the same tests as Lord but using likelihood ratio chi-squares using the Neyman-Pearson lemma. Neyman and Pearson (1928) were able to demonstrate that minus twice the logarithm of a likelihood ratio is asymptotically distributed as χ^2 with the degrees of freedom equal to the difference between the number of free coefficients in the models represented in the numerator and denominator of the ratio. The difference χ^2 used to compare two models is also distributed as a χ^2 with degrees of freedom equal to the difference in degrees of freedom between the two models. Either the chi-square or likelihood-ratio statistic can be used to evaluate model fit for IRT-DIF analyses. However, the likelihood-ratio statistic has an additional desirable property in that it can be partitioned and used to compare the fit of two models.

Thissen, Steinberg, and Wainer (1988) give the following procedure for testing the hypothesis that $b_R = b_F$:

1. The model is fit simultaneously for both groups without any parameter constraints (i.e., $b_R \neq b_F$).

$$G_1^2 = \text{-2(loglikelihood)}$$

is computed for the MML estimates of the item parameters.

2. The same model is fitted except this time $b_R = b_F$ and

$$G_2^2 = \text{-2(loglikelihood)}$$

is computed.

3. The likelihood ratio test is then computed for the significance of the difference between b_R and b_F as

$$G^2(1) = G_2^2 - G_1^2.$$

This will result in one degree of freedom difference between the models. If the difference between the two models is within the limits of chance variation ($\chi^2 < 3.85$, 1 d.f.), then the null hypothesis of no DIF is not rejected. A straightforward extension of this procedure applies when more than one parameter is constrained to be equal across groups such as testlet DIF. The IRT likelihood ratio approach represents a powerful method for investigating the hypothesis of DIF.

Chapter 3

Method and Results

Method

Sample Characteristics

This study used data from an investigation that followed Head Start children into kindergarten (see Bergan, Sladeczek, Schwarz & Smith, 1989). The study was conducted in six different sites: Arizona, California, New Mexico, Iowa, Louisiana, and Mississippi and was comprised of seven school districts and 21 schools. Seven of the schools were rural while the other 14 schools were urban. Rural areas were considered to have less than 2500 inhabitants. The teachers in the study had an average of 11.4 years of teaching experience with an average of 17.6 years of education.

A brief questionnaire was administered to the entire sample (see Bergan et al., 1989) to determine their socioeconomic status (SES). Fifty-three percent of these families responded. Stevens and Featherman's (1981) MESI2 revision of the Duncan Socioeconomic Index was computed for the responding families. The mean for the sample was 27.65. The mean computed for the subsample (n=74) used in this dissertation was 25.70. These results indicate that the children came from primarily economically disadvantaged homes.

The sample consisted of 352 kindergarten children who had complete test and referral data. The mean age for the sample was 67.32 months. The mean age for males (n=186) was 67.97 and was 66.48 months for females (n=166). The sample was composed of the following ethnic groups: (23 %) African American; (45 %) Caucasian; (19 %) Hispanic; (1 %) Asian American; (12 %) Native American. Twenty percent of the sample were referred for special education services. Thirteen percent of the entire sample (n=46) were subsequently placed in special education. Sixteen percent (n=55) of the children were retained. For this data-set, no information was available regarding the criteria used to make referral, placement and retention decisions by the various school districts. Table 1 shows the cells counts and percents for the groupings variables of gender and ethnicity for referral, placement and retention decisions. As might be expected, differences in classification decisions appear to be smaller between the sexes than for ethnicity. DIF for gender would be more unexpected than DIF for ethnicity. The greatest proportional difference is between Caucasian (20.3 %) and non-Caucasian (31.5 %) for referral decisions. If DIF exists it probably lies within this comparison.

Instruments

The dichotomous item responses from various testlets were obtained from a math and reading scale developed for kindergarten age children called the Measurement and Planning System Level K (MAPS-K). Scale construction was guided by a technique developed by Bergan, Stone & Feld (1985). All testlets selected consisted of three items. These are testlets since the items comprising them come from the same domain (i.e., counting objects) and the items comprising them are locally dependent. The testlets from the math

Table 1.

Cell counts and percents using gender and ethnicity for referral, placement, and retention decisions

Grouping variable	Gender		Ethnicity	
	Males (n=186)	Females (n=166)	Caucasian (n=160)	Non-Caucasian (n=192)
Referral				
No	145	134	133	146
Yes	41	32	27	46
Percent referred	28.2	23.8	20.3	31.5
Placement				
No	158	145	139	164
Yes	28	21	21	28
Percent placed	17.7	14.4	15.1	17.0
Retention				
No	154	143	133	164
Yes	32	23	27	28
Percent retained	20.7	16.0	20.3	17.0

scale included identifying geometric shapes, numeral recognition, and counting. The testlets used from the reading scale included letter identification and causal reasoning in stories. Easy items were selected for the testlets in order to maximize information on the low end of the ability continuum. The discrimination parameters, standard errors and location parameters are listed in Table 2 for the items comprising the testlets. These 15 items were estimated together as one set for the entire sample.

A child information form was developed for the study (Bergan, et al., 1989). Teachers were asked to indicate on this form whether a child had been referred, placed, or retained in grade.

Procedure

Since the IRT model used in these analyzes is assumed to be unidimensional, a factor analysis was carried out to determine whether the classification decisions form a factor. A weighted least squares estimator as implemented in **PRELIS2** (Joreskog & Sorbum, 1993) was used to obtain the asymptotic covariance matrix of covariances and correlations. This approach is used when some or all of the variables are censored or categorical. The correlations are shown in Table 3. The correlation between referral and retention is .646 while the correlation between placement and retention is .547. The polychoric correlation between referral and placement is .998. The magnitude of the coefficient is higher than expected based on the different *n* sizes for referral and placement given in Table 1. The polychoric correlation is not a correlation between a pair of scores rather it is an estimate of what *r* would be if each of two categorical variables were in fact continuous and normally

Table 2.

IRT item parameters for the testlet anchors (N=352) using the two-parameter model

Knowledge area	a-parameter	SE	b-parameter	SE
Identifying geometric shapes				
1. Identifying a square object.	0.58	.30	-4.46	.93
2. Identifying a triangle.	0.57	.21	-2.54	.83
3. Identifying a square shaped object.	0.98	.23	-1.59	.32
Numeral recognition				
1. Recognizing numerals from 1 to 5.	1.14	.29	-1.93	.35
2. Recognizing numerals from 6 to 10.	1.86	.27	-0.46	.10
3. Recognizing numerals 11 to 15.	2.03	.29	-0.25	.09
Letter recognition				
1. Identifying uppercase letters.	1.21	.31	-2.14	.41
2. Identifying lowercase letters.	0.86	.23	-2.02	.45
3. Matching two identical letters.	2.50	.42	-0.98	.10
Counting				
1. Counting to 5	.30	.16	-1.96	.09
2. Counting to 10.	1.69	.28	-.92	.13
3. Counting to 5 from a number > 1.	1.19	.20	-.12	.14
Causal reasoning in stories				
1. Recalling the cause of an event in a story.	0.57	.28	-4.52	2.13
2. Recalling the cause of a character's feelings.	0.70	.19	-2.18	.59
3. Identifying the cause of a character's feelings.	0.21	.14	1.37	.12

Table 3.

Correlation matrix for the classification decisions

	Referral	Placement	Retention
Referral	1.000		
Placement	.998	1.000	
Retention	.646	.547	1.000

distributed. In other words, it is the correlation between two latent variables η and ϵ, which are assumed to have a bivariate normal distribution, that underlie two observed variables x and y. This latent correlation is based on the multinomial distribution of the cell frequencies in the contingency table and can be estimated using maximum likelihood (Olsson, 1979). The polychoric estimates will always be somewhat higher than r and is a biased estimate of the true correlation. Brown and Benedetti (1977) showed that this bias is negligible if no expected cell frequencies are less than five. However, for referral and placement the cell consisting of no referral and placement is always empty. This means that the correlation between referral and placement is a biased estimate. Since this relationship exists between referral and placement, their analysis together necessitates the formation of a testlet (Yen, 1993).

The referral and placement testlet constructed is the summed score of either no referral or placement, just referral and no placement, or both referral and placement. The three test items comprising the testlet "Identifying Letters" were used in the factor analysis since they were used as the designated anchor in the DIF analyses. The correlation matrix for the letter items, the referral and placement testlet and retention can be seen in Table 4. Notice that the magnitude of the correlation between retention and the referral and placement testlet is very similar to the correlations in Table 3 between retention and either referral or placement. The referral and placement testlet, retention and the individual items from the letter testlet were loaded on one factor using **LISREL8** (Joreskog & Sorbum, 1993). This tests the hypothesis whether classification decisions can be placed on the latent

Table 4.

Correlation matrix for the letter testlet, referral and placement testlet and retention

	Letter 1	Letter 2	Letter 3	Referral-placement testlet	Retention
Letter 1	1				
Letter 2	0.6	1			
Letter 3	0.33	0.38	1		
Referral-placement testlet	0.1	0.45	0.12	1	
Retention	0.31	0.12	0.31	0.62	1

metric of ability.

A chi-square of 5.74 (p = .22) with 4 degrees of freedom resulted. This indicates that a one factor model fits the data well and that classification decisions and test items can both be measured on the latent ability metric.

The computer program **MULTILOG 6.0** (Thissen, 1991) was used to estimate the parameters for the anchors items, the classification decisions, estimate item information, and test the hypothesis of DIF for each of the testlets. **MULTILOG 6.0** is an extremely flexible program that handles both dichotomous and polytomous item types, mixed models, multi-group analyses and allows for "LISREL" type constraints on the parameters. The ability to have multiple groups and constraints across groups allows the hypothesis of DIF to be investigated.

The math and reading testlets were used to construct the designated anchor. The designated anchor should be free of any DIF and can consist of one to four items. DIF was examined by using gender and ethnicity as the two grouping types. Males were used as the reference group for the gender DIF analysis. Caucasians were used as the reference group for the DIF analysis for ethnicity while the focal group was all other ethnicities. The relatively small sample size precluded any other investigations of DIF for ethnicity.

For all DIF analyses, models were run that compared the loglikelihoods to determine if the standard deviations for the focal and reference groups differed, significantly. One model allowed the standard deviation for the reference group to be free while the standard deviation was constrained to be one for the focal group. This was compared to a model with both standard deviations constrained to be one which is the MULTILOG default. This results in a difference chi-square with one degree of freedom. None of the standard deviations were found to be significantly different. Hence, none of the DIF analyses were run with the standard deviations free to vary.

Pattern counts were obtained for the various grouping levels in order to obtain a likelihood ratio statistic that could be referenced to the chi-square distribution in order to compare model fit determined by subtracting the number of parameters estimated plus one for sample size from the number of pattern counts greater than zero (Thissen, Steinberg, & Wainer, 1988). Most of the analyses took just a few minutes on a 486 personal computer. The number of EM cycles was set high enough to ensure convergence. Most of the runs converged in less than 200 cycles. However, some of the DIF runs using the referral, placement and retention testlet required 400 cycles.

Results

Parameter Estimation

Parameters from a two-parameter model for the various classification decisions are listed in Table 5 calibrated using all five testlets. The discrimination parameter is relatively high and is almost the same for all three decisions types. This indicates that the decisions are strongly related to a single underlying variable. As expected, the location parameter for all three decisions is relatively low on the ability continuum. Notice that the location parameter for placement is lower than the value for referral. Therefore, children who are placed are generally lower on the ability continuum than children who were just referred. The value of the location parameter for retention is similar to the value for placement. Decisions regarding placement and retention occur at approximately at the same place on the ability continuum.

Table 6 shows the parameter estimates for the referral and placement testlet calibrated using all five testlets. The *a* parameter is relatively high while the location parameters occur relatively low on the latent continuum. The first location parameter, b_1, is the point of the ability continuum where the decision has a 50 % chance of being in the referred category or higher. The value of -1.85 indicates that this occurs relatively low on the ability continuum. The b_2 parameter of -1.39 is the point on the ability continuum where the decision has a 50 % chance of being in the no referral or placement category. This value is a little bit more negative than the value for referral

Table 5.

IRT item parameters for the classification decisions using the two-parameter model

Parameters	a-parameter	SE	b-parameter	SE
Referral	1.73	.30	-1.18	0.15
Placement	1.56	.34	-1.64	0.23
Retention	1.60	.28	-1.51	0.2

Table 6.

IRT item parameters for the referral and placement testlet using the graded response model

Parameters	Estimates	SE
a	1.25	.27
b_1	-1.85	.32
b_2	-1.39	.24

given by the two-parameter model. Table 7 shows the parameter estimates for the referral, placement and retention testlet calibrated using all five testlets. Similar sorts of interpretations can be given to the parameters of this testlet. The value of -2.72 for the location parameter b_1 is the point on the ability continuum where the decision has a 50 % chance of being either in the referral and placement category or the referral and retention category or higher. The b_1 parameter occurs very low on the ability continuum. The transition to just referral or retention occurs at -1.63 and the transition to no referral, placement or retention occurs at -1.01.

Item Information

Item information for the three types of classification decisions using the two-parameter model and the two different testlets using the graded response model is given in Table 8 for various levels of θ. The decisions and testlets were calibrated using all five testlets using the entire sample in three separate runs. An examination of the table revels that item information is high for low values of θ as necessitated by the location parameters. Item information is highest in the -2.0 to -1.0 range. The item information is lower for the testlets when compared to the individual items. The item information given for the individual decisions overstates information due to the lack of local independence among the referral and retention (Yen, 1993). Therefore the information give by the testlets more accurately reflects information than the individual decisions.

Table 7.

IRT item parameters for the referral, placement, and retention testlet using the graded response model

Parameters	Estimates	SE
a	1.22	.22
b_1	-2.72	.43
b_2	-1.63	.25
b_3	-1.01	0.18

Table 8.

Item information for classification decisions using the two-parameter model and for two testlets using the graded response model

Theta level	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5
Referral	0.47	0.69	0.73	0.54	0.31	0.15	0.07	0.03
Placement	0.56	0.60	0.48	0.30	0.16	0.08	0.04	0.02
Retention	0.55	0.64	0.55	0.36	0.19	0.10	0.04	0.02
Referral & Placement	0.42	0.43	0.39	0.3	0.2	0.12	0.07	0.04
Referral, Placement, & Retention	0.46	0.45	0.42	0.35	0.26	0.18	0.11	.06

Differential Item and Testlet Functioning

The first step in the DIF analysis was to find a designated anchor. A Rasch model was used that constrained the difficulty parameters to be equal across groups for an item. To ease the burden of finding an anchor, all the items' difficulty parameters were constrained simultaneously to be equal across the focal and reference groups for each of the testlets. The result of these runs were compared to models that allowed the difficulty parameters to vary. This resulted in three degrees of freedom difference between the models. Tables 9 and 10 show the DIF analysis for the testlets: Identifying Geometric Shapes; Numeral Recognition; Counting; Causal Reasoning in Stories; and Letter Identification for both types of grouping variables. A difference chi-square was computed for each model comparison. The hypothesis of no DIF was not rejected for any of the testlets in Tables 9 or 10. The chi-squares for both the "Identifying Numerals" and "Counting" testlets show that they do not fit the data when a Rasch model is used due to unequal slopes. All these testlets could be used for the subsequent DIF analyzes as the designated anchor. However, Identifying Letters was selected as the designated anchor since it showed the least DIF and did not produce sparse cell counts, as mentioned previously. DIF was also investigated for the letter testlet using the two-parameter model for the grouping variables gender and ethnicity. No DIF was detected for either of these comparisons.

Both the Rasch and two-parameter DIF models used all three of the letter items (i.e., Letter 1, Letter 2, and Letter 3). The total number of possible

Table 9.

DIF analysis for gender using five testlets and the Rasch model

Testlet	Model type	df	χ^2	$\Delta\chi$	p
Identifying geometric shapes	$b_1=b_4, b_2=b_5, b_3=b_6$	10	7.6	1.9,df=3	>.50
	free	7	5.7		
Identifying numerals	$b_1=b_4, b_2=b_5, b_3=b_6$	10	19.2*	.2,df=3	>.975
	free	7	19.0*		
Counting	$b_1=b_4, b_2=b_5, b_3=b_6$	10	14.6	1.5,df=3	>.50
	free	7	13.1		
Causal reasoning in stories	$b_1=b_4, b_2=b_5, b_3=b_6$	10	15.6	.7,df=3	>.75
	free	7	14.9		
Identifying letters	$b_1=b_4, b_2=b_5, b_3=b_6$	10	10.2	.3	>.90
	free	7	9.9		

Note:* indicates $p < .05$

Table 10.

DIF analysis for ethnicity using five testlets and the Rasch model

Testlet	Model type	df	χ^2	$\Delta\chi$	p
Identifying geometric shapes	$b_1=b_4,b_2=b_5,b_3=b_6$	10	4.8	1.9,df=3	>.50
	free	7	2.9		
Identifying numerals	$b_1=b_4,b_2=b_5,b_3=b_6$	10	25.3*	0.5,df=3	>.90
	free	7	24.8*		
Counting	$b_1=b_4,b_2=b_5,b_3=b_6$	10	20.3*	1.2,df=3	>.75
	free	7	19.1*		
Causal reasoning in stories	$b_1=b_4,b_2=b_5,b_3=b_6$	10	14.3	1.2,df=3	>.75
	free	7	13.1		
Identifying letters	$b_1=b_4,b_2=b_5,b_3=b_6$	10	13.7	0.6,df=3	>.75
	free	7	13.1		

Note: * indicates p < .05

cell counts was 32, 2^4 for the reference group and 2^4 for the focal group. However, most of these runs had 30 or 28 cells greater than zero. For the referral and placement testlet, just Letter 2 and Letter 3 were used as anchors. If the whole letter testlet had been used, this would have resulted in too many empty cells due to the relatively small sample size. In addition, Letter 2 and Letter 3 were selected since they differed in difficulty, allowing for fewer empty cells. This resulted in total of 24 possible cells, $2 \cdot 2 \cdot 3$ for the reference and $2 \cdot 2 \cdot 3$ for the focal group. This resulted in 22 patterns greater than zero for the ethnicity DIF runs and 24 patterns greater than zero for the gender DIF runs. Letter 2 and Letter 3 were also selected as the anchor for the referral, placement and retention testlet. This resulted in 32 possible cell counts, $2 \cdot 2 \cdot 4$ for the reference and $2 \cdot 2 \cdot 4$ for the focal group. Both the gender and ethnicity DIF runs had full cell counts of 32 for this testlet. The degrees of freedom are calculated by subtracting the number of estimated parameters and the group totals from the total number of cell counts greater than zero.

Table 11 shows the DIF analysis for referral decisions using a Rasch model using the identifying letters testlet as the anchor. Model M1 constrained the location parameters for referral and the anchor items to be equal across males and females. M2 allowed the location parameters to be free across males and females for the referral decision while still constraining the location parameters for the anchor items. A difference chi-square was computed to test the hypothesis of DIF. Failure to reject M1 means that the null hypothesis of no DIF is not rejected. A difference chi^2 was performed to test this hypothesis by subtracting the degrees of freedom and chi-square of M2 from M1.

Table 11.

DIF analysis for referral decisions using the Rasch model

Grouping variable	Model type	df	χ^2	$\Delta\chi^2$	p
Gender					
Model: M1	$b_4=b_8$	25	33.2	.9, df=1	>.25
Model: M2	free	24	32.3		
Ethnicity					
Model: M3	$b_4=b_8$	25	43.8*	1.7, df=1	>.10
Model: M4	free	24	42.1*		

Note: * indicates $p < .05$

Table 11 shows that a chi-square of .9 results with one degree of freedom and a p value of greater than .25. Since M1 was not rejected in favor of M2, the hypothesis of no DIF was not rejected. A similar result was achieved for referral for the grouping variable ethnicity. However, in this case neither of these models fit the data. The difference chi-square that resulted when M3 and M4 are compared failed to reject the hypothesis of no DIF. A Mantel-Haenszel chi-square was also computed to test this hypothesis. The resulting chi-square was 2.26 with one degree of freedom indicating that no DIF exists according to the Mantel-Haenszel statistic. The Mantel-Haenszel is in agreement with the Rasch model. Recall that Table 1 showed the greatest proportional difference is between Caucasian and non-Caucasian for referral decisions. If DIF exists it probably lies within this comparison. However, neither the Rasch model nor the Mantel-Haenszel models detected any DIF.

Similar results were found for placement and referral decisions shown in Tables 12 and 13. The hypothesis of no DIF was not rejected when the Rasch model was used. None of the Rasch models fit for placement decisions. Therefore, the difference chi-squares are at best approximate. Notice that the constrained and unconstrained models fit the same for models M7 and M8.

A different picture emerges when a two-parameter model was used to analyze DIF. Tables 14, 15 and 16 show the results of the DIF analysis for classification decisions using the two-parameter model for the decisions. A two-parameter model was also used for the anchor. Both the location and discrimination parameters were constrained across groups. Constraining the

Table 12.

DIF analysis for placement decisions using the Rasch model

Grouping variable	Model type	df	χ^2	$\Delta\chi^2$	p
Gender					
Model: M5	$b_4=b_8$	21	35.5*	.6, df=1	>.25
Model: M6	free	20	34.9*		
Ethnicity					
Model: M7	$b_4=b_8$	21	34.8*	0, df=1	>.995
Model: M8	free	20	34.8*		

Note: * denotes p < .05

Table 13.

DIF analysis for retention decisions using the Rasch model

Grouping variable	Model type	df	χ^2	$\Delta\chi^2$	p
Gender					
Model: M9	$b_4=b_8$	23	31.2	.4, df=1	>.50
Model: M10	free	22	30.8		
Ethnicity					
Model: M11	$b_4=b_8$	23	29.4	.8, df=1	>.25
Model: M12	free	22	28.6		

Table 14.

DIF analysis for referral decisions using the two-parameter model

Grouping variable	Model type	df	χ^2	$\Delta\chi^2$	p
Gender					
Model: T1	$a_4=a_8$, $b_4=b_8$	22	17.5	1.3, df=2	>.50
Model: T2	free	20	16.2		
Ethnicity					
Model: T3	$a_4=a_8$, $b_4=b_8$	22	27.7	7.6, df=2	<.025
Model: T4	free	20	20.1		

Table 15.

DIF analysis for placement decisions using the two-parameter model

Grouping variable	Model type	df	χ^2	$\Delta\chi^2$	p
Gender					
Model: T5	$a_4=a_8$, $b_4=b_8$	18	19.1	2.4, df=2	>.25
Model: T6	free	16	16.7		
Ethnicity					
Model: T7	$a_4=a_8$, $b_4=b_8$	18	17.1	.7, df=2	>.50
Model: T8	free	16	16.4		

Table 16.

DIF analysis for retention decisions using the two-parameter model

Grouping variable	Model type	df	χ^2	$\Delta\chi^2$	p
Gender					
Model: T9	a_4=a_8, b_4=b_8	20	21.2	.1, df=2	>.95
Model: T10	free	18	21.1		
Ethnicity					
Model: T11	a_4=a_8, b_4=b_8	20	18.8	.7, df=2	>.50
Model: T12	free	18	18.1		

a parameters and *b* parameters to be equal across groups for the decision results in two degrees of freedom difference between the models. The null hypothesis of no DIF was not rejected for gender for referral decisions shown in Table 14. However, the null hypothesis was rejected for ethnicity when referral decisions were analyzed. Table 14 shows that model T3 was rejected in favor of the alternative hypothesis of DIF given by model T4. The difference chi-square was 7.6 with two degrees of freedom and a *p*-value less than .025. The null hypothesis was not rejected for either placement or retention decisions shown in Tables 15 and 16. All the models fit the data unlike the Rasch models given previously.

Table 17 shows the parameter estimates for model T4. As can be seen in Table 17, the *a* parameter is quite different for the reference and focal groups. The *a* parameter is much lower for the focal group indicating a lack of discrimination for referral decisions for ethnicities other than Caucasian. The *b* parameters are also very different for the two groups. The *b* parameter is very low for the focal group and has a very large standard error associated with it. This indicates that referral decisions for ethnicities other than Caucasian are more error prone. This result did not emerge when DIF was analyzed using the Rasch model or the Mantel-Haenszel statistic. Figure 1 shows the trace lines for model T4 for the reference and focal groups. The trace lines overlap indicating nonuniform DIF.

Differential testlet functioning (DTF) was analyzed with the graded response model using the referral and placement testlet and the referral, placement, and retention testlet. Table 18 shows the results of the DTF analysis

Table 17.

Parameter estimates for model (T4) showing DIF for referral decisions

Parameters	a-parameter	SE	b-parameter	SE
Reference group	1.58	0.44	-1.43	0.29
Focal group	0.50	0.23	-2.53	1.18

Figure 1. *Trace lines for model (T4) showing DIF for referral decisions.*

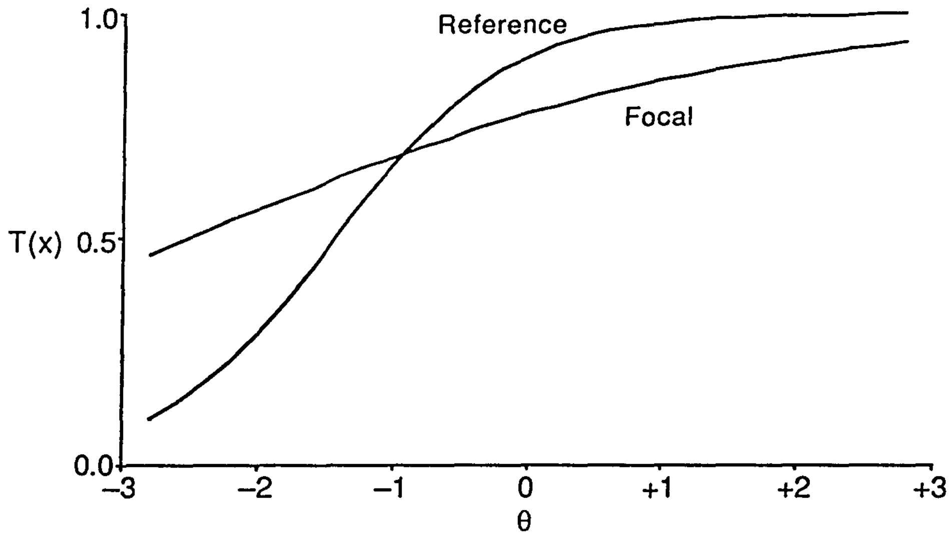


Table 18.

DIF analysis for the referral and placement testlet using the graded response model

Grouping variable	Model type	df	χ^2	$\Delta\chi^2$	p
Gender					
Model: G1	$a_4b_{41}b_{42}=$ $a_8b_{81}b_{82}$	15	7.3	1.7, df=3	>.50
Model: G2	free	12	5.6		
Ethnicity					
Model: G3	$a_4b_{41}b_{42}=$ $a_8b_{81}b_{82}$	15	32.6*	7.9, df=3	<.05
Model: G4	free	12	24.7*		
Model: G5	$a_4=a_8$	13	28.1*	3.4, df=1	>.05

Note: * indicates $p < .05$

when the referral-placement process was analyzed using the graded response model. The referral process was modeled, as mentioned previously, by scoring the decisions as no referral or placement, referral but no placement, both referral and placement. This results in three graded categories. Constraining the discrimination parameter and the location parameters to be equal across groups results in three degrees of freedom. As can be seen in Table 17, the difference chi-square for Gl versus G2 is 1.7 with three degrees of freedom and a p value greater than .50. The hypothesis of no DTF was not rejected. This was not the case for ethnicity. The difference χ^2 was 7.9 (df =3) with a p value less than .05. The hypothesis of no DTF was rejected when the restricted model, G3, was compared with the unrestricted model, G4. Since the difference chi-square was so close to the rejection region of 7.8 for three degrees of freedom, DTF cannot reside in both the location and discrimination parameters. Model G5 was used to test this hypothesis by restricting the a parameters to be equal across groups while leaving the b parameters unconstrained. This results in one degree of freedom difference between the models. The null hypothesis of no DTF for the discrimination parameters was not rejected. Therefore, the DTF must be in the location parameters for this comparison. The graded models for ethnicity did not fit the data. Ethnicity for the referral and placement testlet was the only case when a testlet did not fit the data. Clearly, DTF for ethnicity is functioning in a manner unlike the other testlets. These models did not fit the data. As a result, the difference chi-square is approximate since the likelihood ratio statistic operates under the assumption that the more restricted model is valid.

Table 19 shows similar sorts of DTF models where referral, placement and retention were modeled as a polytomous item. The categories were: no referral, placement or retention; referral but no placement or referral; referral and placement but on retention; and referral, placement, and retention. This results in a testlet with four categories. For the gender and ethnic grouping variables the hypothesis of no DIF was not rejected.

Table 20 shows the parameter estimates for G4. The a parameters are highly similar to those obtained in model T4. The a parameter is exactly the same for the focal group for both models T4 and G4. Once again the a parameter is relatively low for the focal group. The difference in the b_1 parameters is also substantial. As was seen in model T4 the standard error associated with the b_1 parameter for the focal group is very high. The b_1 parameter is the point on the ability continuum where the decision has a 50 % chance of being in the just referred category or above. This is comparable to the results given by model T4. The parameters for b_2 also differ although not as substantially. Figure 2 shows the trace lines for model G4. The trace lines for the second and third score categories for the reference and focal groups overlap indicating nonuniform DTF. The conclusion is that DIF and DTF exists for referral decisions for the focal group when ethnicity is the grouping variable.

Table 19.

DIF analysis for the referral, placement, and retention testlet using the graded response model

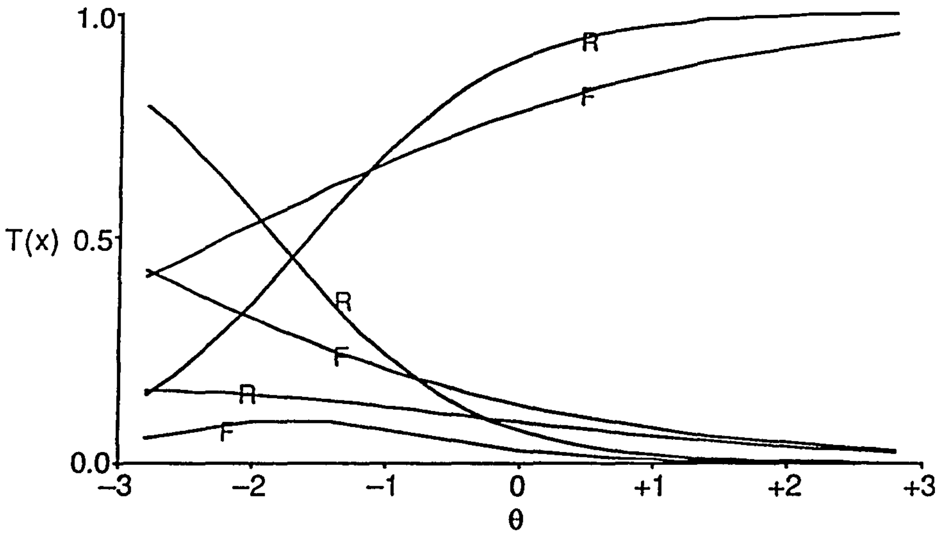
Grouping variable	Model type	df	χ^2	$\Delta\chi^2$	p
Gender					
Model: G6	$a_4b_{41}b_{42}b_{43}=$ $a_8b_{81}b_{82}b_{83}$	22	18.4	1.5 df=4	>.75
Model: G7	free	18	16.9		
Ethnicity					
Model: G8	$a_4b_{41}b_{42}b_{43}=$ $a_8b_{81}b_{82}b_{83}$	22	13.0	2.8,df=4	>.50
Model: G9	free	18	10.2		

Table 20.

Parameter estimates for model (G4) showing DTF for the referral and placement testlet

Parameters	a	SE	b_1	SE	b_2	SE
Reference group	1.39	0.43	-1.84	0.44	-1.57	0.37
Focal group	0.58	0.25	-3.30	1.34	-2.19	0.90

Figure 2. *Trace lines for model (G4) showing DTF for the referral and placement testlet.*



Note: R indicates the reference group and F indicates the focal group.

Chapter 4

Discussion

This dissertation sought to answer a number of questions regarding classification decisions using item response theory that had not been asked previously. One question was whether classification decisions could be empirically placed on the latent continuum of ability usually associated with test items. The resulting "item" parameters were similar to those that might be expected from conventional test items. When the a parameters were estimated for the entire sample, they were found to be uniformly high for both the individual classification decisions and for the testlets. This indicates that the decisions are strongly related to a single underlying variable. This conclusion is consistent with the results of the confirmatory factor analysis indicating an acceptable fit for the letter testlet, the referral-placement testlet and retention decisions. Obtaining high a parameters also indicates that there is adequate discrimination for the various classifications as a function of ability. A high positive value for the a parameter indicates that the classification decisions discriminate well between children who were either referred or not referred, placed or not placed, and retained or not retained. The estimates obtained for the location parameters for the various classification decisions were all relatively low on the ability continuum, as expected. The location parameter for placement was lower on the ability continuum than for referral. This is reasonable since it is assumed that children who were subsequently placed in special education are lower on the ability continuum than children who were just referred. The value for retention was close to the value for placement. It is suspected that for a sample with a smaller percentage of referred, placed, and retained children, the *b* parameters would have been much lower on the ability continuum than the ones obtained here.

Analyzing referral and retention decisions is more problematic than for most item types due to their complete dependency. If a child is placed then the child must have been previously referred. If a two-by-two cross-tabulation is created consisting of placed and not placed and referred and not referred then one of the cells is always empty. This empty cell is the not-referred/placed cell. The referral by placement table forms a perfect Guttman scale of 00, 10 and 11 since the 01 cell cannot occur. This necessitates the creation of a testlet. Two testlets were created; a referral and placement testlet and a referral, placement, and retention testlet. Since these are polytomous variables, the graded response model was used to analyze them. Both testlets were found be graded. The first location parameter for the score category of both referral and placement testlet was lower than any of the location parameters for the individual decisions. Similarly, the first location parameter for the score category of referral, placement and retention was much lower on the ability continuum than just referral and placement or referral and retention for the referral, placement, and retention testlet.

Testlet formation is not only useful for analyzing classification decisions such as referral, placement, or retention in grade but also could be employed in the analysis for other types of decisions. Many decision-making processes might conceivably consist of four or five decisions that are all dependent. For example, in a clinical situation we might have a test which is used to make a decision, then a treatment is administered, then another decision is made based on the outcome of the treatment and so on. This decision process might be best analyzed using a testlet given that many decisions could be dependent upon previous ones. If the decision process was not ordered, then Bock' (1972) nominal response model could be used to "score" the testlet.

Obtaining item information for classification decisions was another focus of this dissertation. Item information was slightly higher for the individual decisions as opposed to the values for the testlets. Likewise, the marginal reliability for the three decisions scored individually was .74 while marginal information was .69 for the referral and placement testlet and was .70 for the referral, placement, and referral testlet. [1] Little information is lost when decisions are scored as a testlet as opposed to being scored individually. Given these results, it appears that the best course might be to score the decisions individually since the standard errors are lower and consequently item information is higher. However, as Thissen, Steinberg, and Mooney (1989) have pointed out for other cases, the information computed for the three decisions is under the assumption that the three decisions are independent of each other. This is not the case since referral and placement are completely de-

[1] Since error varies as a function of ability, it cannot be summarized by single a indicator of reliability (Shermis, M. 1992). However, a marginal reliability coefficient can be calculated by averaging the marginal error variance, $\overline{\sigma_e}$ across the distribution of ability where $\overline{\sigma_e} = \int \overline{\sigma_e} g(\theta) d(\theta)$. The marginal reliability coefficient is given as

$$\overline{p} = \frac{\sigma_\theta^2 - \overline{\sigma_e}}{\sigma_\theta^2}$$

where σ_θ^2 is the variance of the estimates.

pendent. Therefore, the information computed when the decisions are scored individually is not completely correct. Information is over stated for the case of the individual classification decisions. Therefore, the information from the testlets is at least as valid, if not more so, than when the decisions are scored individually. This suggests that testlet formation is a very useful tool when "items" are not independent.

To date, item response theory has been, as the name suggests, concerned with the processes underlying an examinee's response to a test item and hence individual differences. Item response theory methodology has matured to the extent, like many other statistical models, that it should be applied to a broader class of problems than just test items. One purpose of this dissertation was to redefine what constitutes a test item in order to explore new questions. Analyzing classification decisions such as referral, placement and retention constituted one such question. As alluded to previously, item response methodology might also be extended to the analysis of other classification decisions. For example, IRT methodology could be used to analyze selection or placement decisions. Thissen (1993) suggests that items do not necessarily have to be questions and gives one example where he employed the graded response model to analyze number of mental health admissions and age at first admission as predictors of violence in adult male mental health admissions (Klassen & O'Connor, 1987). Many such potential applications could be conjectured that allow for new questions to be addressed.

A second question pertained to the existence of DIF and DTF for classification decisions using gender and ethnicity as grouping variables and whether

DIF would emerge at a more "macro" level of analysis. The possibility of DIF was suspected based on the proportional differences in referral for the ethnicity grouping variable. A Mantel-Haenszel (MH) chi-square was computed to investigate this hypothesis. It failed to reveal any DIF, as did any of the Rasch models. The trouble with the MH statistic, as Bock (1993) suggests, is that it only shows the difference in the intercepts of the response functions in the focal and reference groups. It provides no information whether the functions have different slopes. As a result, it does not reveal what portion of the ability distribution is affected by DIF. The same types of criticism could be attributed to the Rasch model. A different picture did emerge when the individual decisions were analyzed for DIF using the two-parameter model. There was a substantial difference in the *a* parameters for the two groups for ethnicity when referral was analyzed. The *a* parameter for the reference was three times larger for the reference group when compared to the focal group. The location parameters for the two groups were also substantially different with a large standard error for the focal group. None of the other comparisons revealed any DIF. Using the two-parameter model as well as the graded models, with such a small sample size (N=352) that is split between two groups, could have lead to a lack of power. The parameter estimates had to be very different before the null hypothesis was rejected.

The differential testlet functioning (DTF) model comparisons were similar, in part, to those obtained from the two-parameter model. The models fit for the gender comparison and there was no DTF. The referral and placement testlet showed DTF for ethnicity. Adding placement to referral tended to reduce the DTF since there was no DIF for placement. However, the results differed in that DTF resided in the location parameters as opposed to the discrimination parameters. However, none of the graded models fit for when ethnicity was the grouping variable.

No DTF emerged for the referral, placement, and retention testlet. Adding retention to the referral and placement testlet appears to eliminate the DTF shown by the referral and placement testlet. Since there was very little difference between the groups for retention decisions, adding it to referral and placement was enough to eliminate the DTF shown by the referral and placement testlet. Obviously, the "items" used to construct a testlet will determine if DTF is present.

However, much greater weight should be given to modeling referral and placement since this is an important process occurring in schools. Referral decisions for ethnicity (Caucasian verses non-Caucasian) differ as function of ability. Teachers in this limited sample tended to make referral decisions that did not take ability into account for this comparison. Bergan et al. (1989) showed that a child's ability determined whether he or she would be referred, placed or retained. Children with high math and reading scores were unlikely to be referred, placed or retained. and were less likely to need these services. Fortunately, there was no DIF with regard to placement. Apparently, the interdisciplinary evaluation team used to make the placement decision does take ability into account when making the decision. Assessment data collected by the school psychologist is being taken into account in determining whether a child is placed.

Clearly, referral functions dissimilarly in different ethnic groups. If only the MH statistic or the Rasch model been used, no DIF would have been detected. Therefore, DIF emerged at a more "macro" level of analysis with the two-parameter model and the referral-placement testlet using the graded response model. Failure to check for differences with regard to the slope parameter disregards an important source of DIF and DTF for classification decisions.

DIF (or DTF) is performed to ensure that test scores for various groups are comparable. This is important since tests perform a variety of important societal functions such as assignment of patients to therapeutic treatments, vocational guidance, certification, and the promotion of individuals. If a test could be shown to be free of DIF, it could then be assumed to lead to equitable treatment. What this dissertation has done is to take DIF a step further and analyze the decision themselves. Messick (1989) suggests that the social consequences stemming from a test are an important component to validity. Analyzing classification decisions stemming from a test for DIF supplies support for this important validity component. Van der Linden (1991) states that test use has it origins in the necessity for selection and placement decisions in education, the army, and public administration. He gives the example of Binet's pioneering test development used for the assignment of retarded children to special education. Although testing practice has roots in decision making, it has evolved mainly as a theory of measurements (i.e., ability estimation). Cronbach's and Gleser's (1965) *Psychological Tests and Personnel Decisions* is the only modern treatise attempting to provide test-based decision making with a sound theoretical basis. If tests are used to make decisions, then we should gather information as to the validity of those decisions. Item response theory can help provide this important validity information regarding decision making.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andersen, E.B. (1972). The numerical solution to a set of conditional estimation equations. *Journal of the Royal Statistical Society, 26*, 42-54.

Andersen, E. B. (1988). Comparison of latent structure models In R. Langeheine & J. Rost (Eds.) *Latent trait and latent class models* (pp.207-228) New York: Plenum.

Angoff, W. H. (1993). Perspectives on differential item functioning In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24) Hillsdale, NJ:Lawerance Erlbaum Associates.

Bergan J. R., Sladeczek, I. Schwarz, R. D., & Smith, A. (1989). Effects of a measurement and planning system on kindergartner's cognitive development and educational programming. *American Educational Research Journal 28*, 683-714.

Bergan, J.R., Stone C.A., & Feld J.K. (1985). Path-referenced assessment of individual differences. In C.R. Reynolds & V.L. Wilson (Eds.) *Methodological and statistical advances in the study of individual differences.* New York: Plenum Press.

Bartholomew, D.J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society, B, 42*, 293-321.

Birnbaum, A. (1968). In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bock, (1993). Different DIFs: Comment on the Papers Read by Neil Dorans and David Thissen In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 115-122) Hillsdale, NJ:Lawerance Erbaum Associates.

Bock, R.D., & Liberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 46*, 443-449.

Bock, R.D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443-449.

Brown, M.B., & Bendetti, J.K. (1977). On the mean and the variance of the tetrachoric correlation coefficient. *Psychometrika, 42*, 347-355.

Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 44*, 5-32.

Cronbach, L. J., & Gleser, G.C.(1965). *Psychological Tests and Personnel Decisions* 2nd ed. Urbana, IL:University of Illinois Press.

Demspter, A.P. Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm with discussion. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

Dorans, N.J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Dorans, N.J., & Holland, P.W.(1993). DIF detection and description: Mantel-Haenszel and Standardization In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24) Hillsdale, NJ:Lawerance Erbaum Associates.

Fisher, R.A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver and Boyd.

Joreskog, K.J., & Sorbum, D. (1993). *LISREL8: Analysis of linear structural equations.* Mooresville, IN:Scientific Software, Inc.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145) Hillsdale, NJ:Lawerance Erbaum Associates.

Kelderman, H., & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27*, 307-328.

Klassen, D., & O'Connor, W. A. (1987, October). *Predicting violence in mental patients: Cross-validation of an actuarial scale.* Paper presented the annual meeting of the American Public Health Association,.New Orleans.

Lawley, D.N. (1943). On problems connected with item selection and test construction. *Recordings of the Royal Society of Edinburgh, 61*, 273-287.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monographs (Whole No. 7)*, Richmond, VA: William Byrd Press.

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

Lord, F.M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from the retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Masters, G.A. (1982). Rasch model for the partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, (1988). Measurement models for ordered response categories. In R. Langeheine & J. Rost (Eds.) *Latent trait and latent class models* (pp. 11-28) New York: Plenum.

Messick, S. (1989). Validity. In R. Linn (Ed.) *Educational Measurement* (pp. 13-104) New York, NY: Macmillan.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551-560.

Muthen, B. (1979). A structural probit model with latent variables. *Journal of the American Statistical Association, 74*, 807-811.

Muthen, B., & Christofferson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46*, 407-419.

Neyman, J., & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20A*, 174-240, 263-294.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460.

Rasch, G. (1960). *Probabilistic models for some intelligence and attaintment tests.* Copenhagen: Denmarks Paedagogiske Institute. (Republished in 1980 by the University of Chicago Press, Chicago).

Rasch, G. (1966). An item analysis which takes individuals differences into account. *British Journal of Mathematical and Statistical Psychology, 19*, 49-57.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reschly, D.J. (1988). Minority MMR Overrepresentation and Special Education Reform. *Exceptional Children, 54*, 316-323.

Richardson, M.W. (1936). The relation between difficulty and the differential validity of a test. *Psychometrika, 1*, 33-49.

Rosenbaum, P.R. (1988). Item bundles. *Psychometrika, 53*, 349-359.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores, *Psychometric Monograph No. 17* Richmond, VA:William Byrd Press.

Shepard, L.A. (1982). Definitions of bias. In R.A. Berk (Ed.) *Handbook of methods for detecting test bias* (pp. 9-30). Baltimore: John Hopkins University Press.

Shepard, L.A. (1989). Identification of mild handicaps. In R.L. Linn (Ed.) *Educational Measurement* (pp. 545-572) New York: ACE/Macmillian.

Shermis, M. (1992, April). *Assessing the reliability of computer adaptive testing branching algorithms using hyperCAT*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Stevens, G., & Featherman, D. (1981). A revised socioeconomic index of occupational Status. *Social Science Research, 10*, 364-395.

Stroud, A.H., & Sechrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs (N.J.): Prentice-Hall.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one parameter logistic model. *Psychometrika, 47*, 201-214.

Thissen, (1991). *MULTILOG user's guide (Version 6)* Mooresville, IN Scientific Software.

Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. In N. Frederiksen, R. Mislevy, & I. Bejar (Eds.) *Test theory for a new generation of tests.* (pp. 79-98) Hillsdale, NJ:Lawerance Erbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H.(1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-114) Hillsdale, NJ:Lawerance Erbaum Associates.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*, 385-395.

Thissen, D., Steinberg, L., & Gerrad, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in traces lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Mooney, J.A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247-260.

Van der Linden, W. (1991). Applications of decision theory to test-based decision making. In R. Hambleton & J. Kaal (Eds.) *Advances in educational and psychological testing* (pp. 129-156) Norwell, MA:Kluwer Academic Publishers

Wainer, H., & Kiley, G.L. (1987). Item clusters and computer adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H., Sireci, S.G., & Thissen, D. (1991). Differential Testlet Functioning: Definitions and Detection *Journal of Educational Measurement, 26*, 197-219.

Wainer, H. (1985). On the study of matching cut-scores to test characteristics: An observed score approach. Technical Report No. 85-86 Educational Testing Service.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-214.