ASSESSING UNIDIMENSIONALITY OF TEST ITEMS AND SOME ASYMPTOTICS OF PARAMETRIC ITEM RESPONSE THEORY

BY

AMY GOODWIN FROELICH

B.S., University of Illinois at Urbana-Champaign, 1994

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2000

Urbana, Illinois

ASSESSING UNIDIMENSIONALITY OF TEST ITEMS AND SOME ASYMPTOTICS OF PARAMETRIC ITEM RESPONSE THEORY

Amy Goodwin Froelich, Ph.D.
Department of Statistics
University of Illinois at Urbana-Champaign, 2000
William Stout, Advisor

In the field of item response theory (IRT), the DIMTEST procedure (Stout, 1987) and the Poly-DIMTEST procedure (Li, 1995) provide a nonparametric hypothesis test of unidimensionality for a given test data set. A new bias correction method for the DIMTEST and Poly-DIMTEST procedures is developed based on the nonparametric IRT bootstrap method (Kim, 1994). Using this method, simulations show the new procedures have Type I error rates near the nominal rate of rejection and good power to detect multidimensionality present in test data in most simulated situations. By replacing the original bias correction method developed by Stout (1987), the new DIMTEST and Poly-DIMTEST procedures provide the user with greater flexibility and better statistical performance on a large variety of test data.

The new DIMTEST procedure is then applied to test items from a Computer Adaptive Test (CAT). Using a multi-stage testlet design (Armstrong, Jones, Koppel & Pashley, 1999) and the new DIMTEST procedure, the CAT-DIMTEST procedure is developed to test the null hypothesis of unidimensionality between the CAT's pretest items and the CAT's operational items. Simulations show the new CAT-DIMTEST procedure has a Type I error rate slighly below the nominal rate of $\alpha = 0.05$ and good power to detect multidimensionality in most simulated situations.

Using results from He & Shao (2000), a proof of the consistency and asymptotic normality of item parameter estimates obtained from the Marginal Maximum Likelihood Estimation (Bock & Lieberman, 1970) procedure as both the number of examinees and the number of items tends to infinity is presented. The proof depends upon fairly general regularity conditions on the model and on the growth of the number of items relative to the number of examinees.

To my husband and my parents.

Acknowledgments

I wish to express my sincere gratitude to my advisor, Professor William Stout. This thesis would not have been possible without his guidance and support.

I would like to thank the members of my thesis committee, Professor Terry Ackerman, Professor Xuming He, and Professor John Marden, for their helpful comments and suggestions. I am particularly grateful to Professor Xuming He for his guidance and direction on Chapter Five.

I am indebted to both the current and former members of the Statistical Laboratory for Educational and Psychological Measurement for their camaraderie. In particular, I would like to express my appreciation to Professor Brian Habing for his help in editing my thesis and his work on the kernel smoothing procedure used in Chapter Two, to Professor Ratna Nandakumar for providing the factor analysis program used in Chapter Three, and to Dr. Louis Roussos for his insightful comments regarding the research in Chapters Two and Four. Lastly, I would like to express my sincere thanks to my officemate Sarah Hartz for her friendship and support over the past four years.

I would like to thank the faculty and staff of the Department of Statistics for providing a very congenial learning environment. In particular, I would like to thank Professor Barbara Bailey for her never-ending help with formatting in LaTex, and Usha Dhar and Lisa Yanello for their assistance in countless ways.

Finally, my deepest love and appreciation go to my husband, Jim, for his love, support and patience over the years and to my parents, Jim and Kay Goodwin, and my brother Scott, for their constant love and support of all my endeavors.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction to Item Response Theory

The field of Item Response Theory (IRT) is concerned with statistically modeling examinee responses to test questions, or items, based on a latent ability vector, denoted as $\mathbf{\Theta}$. The random latent vector $\mathbf{\Theta}$ can be thought of as the set of dimensions or abilities that affect an examinee's responses to test items. For example, a math test might contain dimensions related to different areas of mathematics, such as algebra, geometry, or trigonometry, while a reading comprehension test might contain an overall reading ability dimension plus dimensions related to the subject of the passages, such as history, science, or current events. Examinee responses to test items can also be affected by other characteristics of the test, such as the format of the items or whether the test is administered with a time limit.

When a test is administered, a response pattern for each examinee is obtained. Let U_i be the response of a randomly sampled examinee on the ith item of a test, and let $\mathbf{U}^{(n)}$ $= (U_1, U_2, \ldots, U_n)^T$ denote the response pattern of a randomly sampled examinee to an n item test. Items can be scored dichotomously, where $U_i = 1$ if the examinee answers item i correctly and $U_i = 0$ if the examinee answers item i incorrectly, or polytomously, where $U_i = \{0, 1, \ldots, r_i\}$ denotes an examinee's score on item i with maximum possible score equal to r_i. The number of possible score categories for item i is then $r_i + 1$.

Examinee responses to test items depend both upon the characteristics of the items themselves and upon the latent ability vector $\boldsymbol{\Theta}$. The probability a randomly sampled examinee with ability vector $\boldsymbol{\Theta} = \boldsymbol{\theta}$ answers item i correctly is given by the conditional probability

$$P_i(\boldsymbol{\theta}) = P(U_i = 1 | \boldsymbol{\Theta} = \boldsymbol{\theta}). \tag{1.1}$$

The model in equation (1.1) is defined as the Item Response Function or IRF for item i. Likewise, for polytomous items, the probability a randomly sampled examinee with ability vector $\boldsymbol{\Theta} = \boldsymbol{\theta}$ obtains at least score q on an item with maximum score r_i is given by the conditional probability

$$P^{\star}_{i,q}(\boldsymbol{\theta}) = P(U_i \geq q | \boldsymbol{\Theta} = \boldsymbol{\theta}). \tag{1.2}$$

The probability a randomly sampled examinee with ability vector $\boldsymbol{\Theta} = \boldsymbol{\theta}$ obtains score q on an item with maximum score r_i is then given by the difference

$$P_{i,q}(\boldsymbol{\theta}) = P(U_i = q | \boldsymbol{\Theta} = \boldsymbol{\theta}) = P^{\star}_{i,q}(\boldsymbol{\theta}) - P^{\star}_{i,q+1}(\boldsymbol{\theta}). \tag{1.3}$$

The model in equation (1.2) is defined as the cumulative Item Category Response Function for score category q of item i and the model in equation (1.3) is defined as the Item Category Response Function (ICRF) for score category q of item i. Finally, the probability a randomly sampled examinee with ability vector $\boldsymbol{\Theta} = \boldsymbol{\theta}$ obtains a particular response pattern $\mathbf{u}^{(n)}$ is given by the conditional probability

$$P(\mathbf{U}^{(n)} = \mathbf{u}^{(n)} | \boldsymbol{\Theta} = \boldsymbol{\theta}). \tag{1.4}$$

For the latent variable models defined in equations (1.1), (1.2) and (1.4) to be psychometrically reasonable, two assumptions are often made. First, the latent variable probability model is assumed to be monotone for each item. A latent variable model for dichotomous item i is **monotone** if equation (1.1) increases coordinatewise in $\boldsymbol{\theta}$. Thus, the probability of a correct response on item i increases as the value of latent examinee ability vector $\boldsymbol{\theta}$

increases. Likewise, a latent variable model for polytomous item i with maximum score r_i is **monotone** if equation (1.2) increases coordinatewise in $\boldsymbol{\theta}$ for all $q = 1, \ldots, r_i$. Thus, the probability of obtaining at least score q on item i increases as the value of the latent examinee ability vector $\boldsymbol{\theta}$ increases.

Second, the latent variable probability model is assumed to be locally independent. A latent variable model is **locally independent** if for all $\boldsymbol{\theta}$ and all possible response patterns $\mathbf{u}^{(n)}$,

$$P(\mathbf{U}^{(n)} = \mathbf{u}^{(n)} | \boldsymbol{\Theta} = \boldsymbol{\theta}) = \prod_{i=1}^{n} P(U_i = u_i | \boldsymbol{\Theta} = \boldsymbol{\theta}). \tag{1.5}$$

Thus, examinee responses to test items are independent conditioned upon the value of the latent examinee ability vector $\boldsymbol{\Theta}$. In other words, the latent ability vector $\boldsymbol{\Theta}$ contains all possible dimensions or abilities that affect examinee performance on the test items.

The dimensionality of a test is defined as the minimum number of dimensions of the latent vector $\boldsymbol{\Theta}$ required to produce a locally independent and monotone latent variable probability model. A distinction is often made between models having dimensionality of one or more than one. If the number of dimensions of the latent ability vector is one, the latent variable model is called **unidimensional** and if the number of dimensions of the latent ability vector is more than one, the latent variable model is called **multidimensional**.

1.1 Parametric Models for Dichotomous Items

Given the assumptions made on the latent variable probability model above, any parametric model used to explain an IRF should be monotone coordinatewise in $\boldsymbol{\theta}$ and have zero and one as its minimum and maximum values respectively. In the early development of IRT, the cumulative density function of the normal distribution, also called the normal ogive, was used to model an item's response function (Lord, 1952). Thus, the probability a randomly sampled examinee with unidimensional ability θ answers item i correctly is determined using the formula,

$$P_i(\theta) = \int_{-\infty}^{\theta} \phi(t)dt = \Phi(\theta), \tag{1.6}$$

where $\phi(\cdot)$ is the density and $\Phi(\cdot)$ is the distribution function of the standard normal distribution. Due to the indetermancy of the normal ogive function, calculations from this model can only be made using a computer or prepared tables.

In place of the normal ogive model, Birnbaum (1968) proposed using a logistic function to model IRFs. The probability a randomly sampled examinee answers item i correctly is determined using the formula,

$$P_i(\theta) = LGT(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \frac{1}{1+\exp(-\theta)}. \tag{1.7}$$

Since calculations from this model can be made directly, the logistic function is commonly used to model IRFs. Haley (1952) found that for all x,

$$\left| \int_{-\infty}^{x} \phi(t)dt - LGT(1.7x) \right| \leq 0.01.$$

Therefore, the constant 1.7 is used to equate the logistic model to the normal ogive model.

Three different logistic models are commonly used to model unidimensional IRFs. The one parameter logistic (1PL) model, also known as the Rasch model, defines the probability of a correct response on item i for a randomly sampled examinee with ability θ as

$$P_i(\theta) = \frac{1}{1+\exp[-1.7(\theta - b_i)]}, \tag{1.8}$$

where b_i is called the item's difficulty parameter. The difficulty parameter b_i is the value of θ when the probability of a correct response on item i is 0.5.

The two parameter logistic (2PL) model defines the probability of a correct response on item i for a randomly sampled examinee with ability θ as

$$P_i(\theta) = \frac{1}{1+\exp[-1.7a_i(\theta - b_i)]}, \tag{1.9}$$

where b_i is the item's difficulty parameter and a_i is called the item's discrimination parameter. Again, the difficulty parameter b_i is the value of θ when the probability of a correct response on item i is 0.5 and the parameter a_i is the slope of the IRF at the point $\theta = b_i$. Thus, the a_i parameter is a measure of how well the item discriminates between the performance of examinees whose abilities are close to b_i.

The three parameter logistic (3PL) model defines the probability of a correct response on item i for a randomly sampled examinee with ability θ as

$$P_i(\theta) = c_i + \frac{1-c_i}{1+\exp[-1.7a_i(\theta - b_i)]}, \tag{1.10}$$

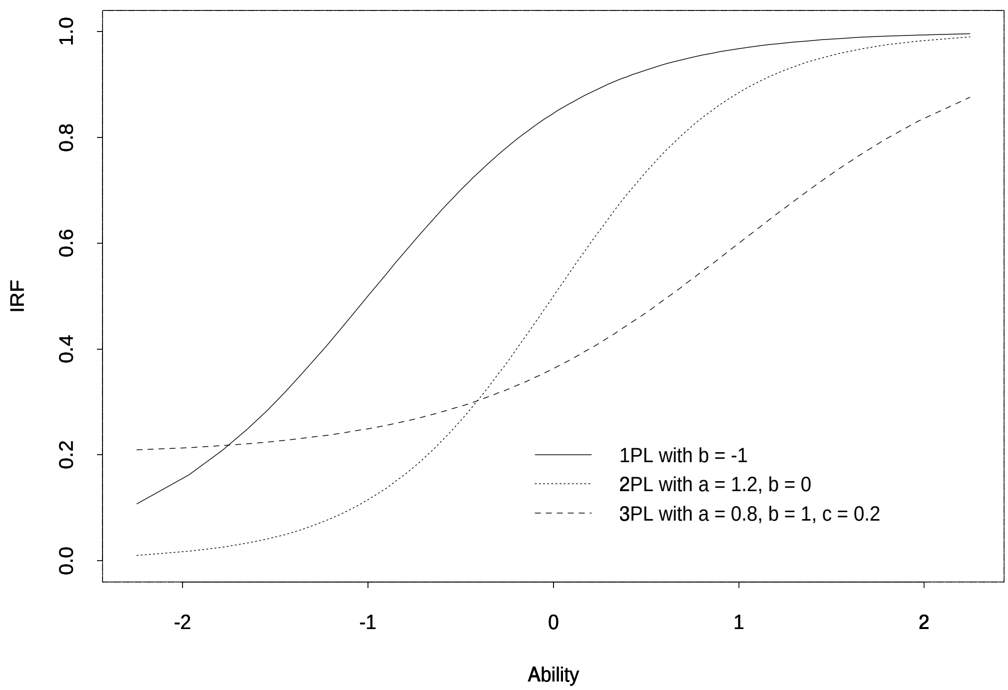
where b_i and a_i are the item's difficulty and discrimination parameters and c_i is called the item's guessing parameter. The parameter c_i is the probability an examinee with ability $\theta = -\infty$ will answer item i correctly. Thus, in some sense, the c_i parameter in the 3PL model estimates the amount of examinee guessing on item i. The difficulty parameter b_i is now the value of θ when the probability of a correct response on item i is $(1+c_i)/2$ and the discrimination parameter a_i is the slope of the IRF at the point $\theta = b_i$. Examples of the 1PL, 2PL and 3PL unidimensional models are given in Figure 1.1.

When the data are multidimensional, a generalization of the above models is used to explain equation (1.1). Let d equal the dimensionality of a test. The probability a randomly sampled examinee with ability vector $\boldsymbol{\theta}$ correctly answers item i is given by

$$P_i(\boldsymbol{\theta}) = c_i + \frac{1-c_i}{1+\exp[-1.7\mathbf{a}_i^T(\boldsymbol{\theta} - \mathbf{b}_i)]}, \tag{1.11}$$

where the vector $\mathbf{a}_i = (a_{i,1}, \ldots, a_{i,d})^T$ consists of discrimination parameters for each dimension for item i, the vector $\mathbf{b}_i = (b_{i,1}, \ldots, b_{i,d})^T$ consists of difficulty parameters for each dimension for item i, and c_i is the guessing parameter for item i.

Figure 1.1: Unidimensional Parametric Models for Dichotomous IRFs



1.2 Parametric Models for Polytomous Items

Given the assumptions made on the latent variable probability model, a parametric model to explain the cumulative ICRF for each score category q should be monotone in $\boldsymbol{\theta}$ and have zero and one as its minimum and maximum values respectively. The three most commonly used models for unidimensional polytomous items are the Graded Response Model (GRM), the Partial Credit Model, and the Generalized Partial Credit Model (GPCM).

The Graded Response Model (Samejima, 1969) defines the cumulative ICRF for item i

for score categories $q = 0, 1, \ldots, r_i, r_i + 1$. For score categories 0 and $r_i + 1$, the values of $P^{\star}_{i,0}(\theta)$ and $P^{\star}_{i,r_i+1}(\theta)$ are defined as 1 and 0 respectively. For score categories $q = 1, \ldots, r_i$, the value of the cumulative ICRF is given by

$$P^{\star}_{i,q}(\theta) = \frac{1}{1 + \exp[-1.7a_i(\theta - b_{i,q})]}. \tag{1.12}$$

Similar to the 2PL model for dichotomous items, the parameter $b_{i,q}$ is the difficulty parameter for score category q of item i and the parameter a_i is the discrimination parameter for item i. The difficulty parameter $b_{i,q}$ is the value of θ when the probability of obtaining at least score q on item i is 0.5 and the discrimination parameter a_i is the common slope of the ICRFs at the point $\theta = b_{i,q}$. Figure 1.2 contains an example of the cumulative ICRFs from the GRM for an item with five categories. The values of the parameters used in the figure are $a_i = 1.2$ and $b_{i,q} = (-1, -0.25, 0.25, 1)$ for $q = 1, \ldots, 4$.

Instead of defining the cumulative ICRF, the Partial Credit Model (Masters, 1982) defines the probability a randomly sampled examinee obtains score q (the ICRF) on item i as

$$P_{i,q}(\theta) = \frac{\exp\left[\sum_{v=1}^{q} 1.7(\theta - b_{i,v})\right]}{1 + \sum_{f=1}^{r_i} \exp\left[\sum_{v=1}^{f} 1.7(\theta - b_{i,v})\right]}, \tag{1.13}$$

for $q = 1, \ldots, r_i$. For score category $q = 0$, the item category response function is defined as

$$P_{i,0}(\theta) = 1 - \sum_{v=1}^{r_i} P_{i,v}(\theta). \tag{1.14}$$

Similar to the Rasch model for dichotomous items, the Partial Credit model includes only a difficulty parameter $b_{i,q}$ for score category q of item i but forces all items to have the same discrimination.

Muraki (1992) generalized the Partial Credit Model to include an item discrimination parameter for each item. Thus, the Generalized Partial Credit Model (GPCM) defines the

ICRF for item i as

$$P_{i,q}(\theta) = \frac{\exp\left[\sum_{v=1}^{q} 1.7a_i(\theta - b_{i,v})\right]}{1 + \sum_{f=1}^{r_i} \exp\left[\sum_{v=1}^{f} 1.7a_i(\theta - b_{i,v})\right]}, \qquad (1.15)$$

for $q = (1, \ldots, r_i)$. For score category $q = 0$, the item category response function is defined as

$$P_{i,0}(\theta) = 1 - \sum_{v=1}^{r_i} P_{i,v}(\theta). \qquad (1.16)$$

The parameter $b_{i,q}$ is a category parameter for score category q of item i and the parameter a_i is a common discrimination parameter for item i. The $b_{i,q}$ parameter is defined as the value of θ where the ICRFs $P_{i,q-1}(\theta)$ and $P_{i,q}(\theta)$ intersect. Figure 1.3 contains an example of the ICRFs from the GPCM for an item with five categories. The values of the parameters used in the figure are $a_i = 1.2$ and $b_{i,q} = (-1.5, -0.5, 0.5, 1.25)$.

Figure 1.2: Graded Response Model for Unidimensional Polytomous ICRFs

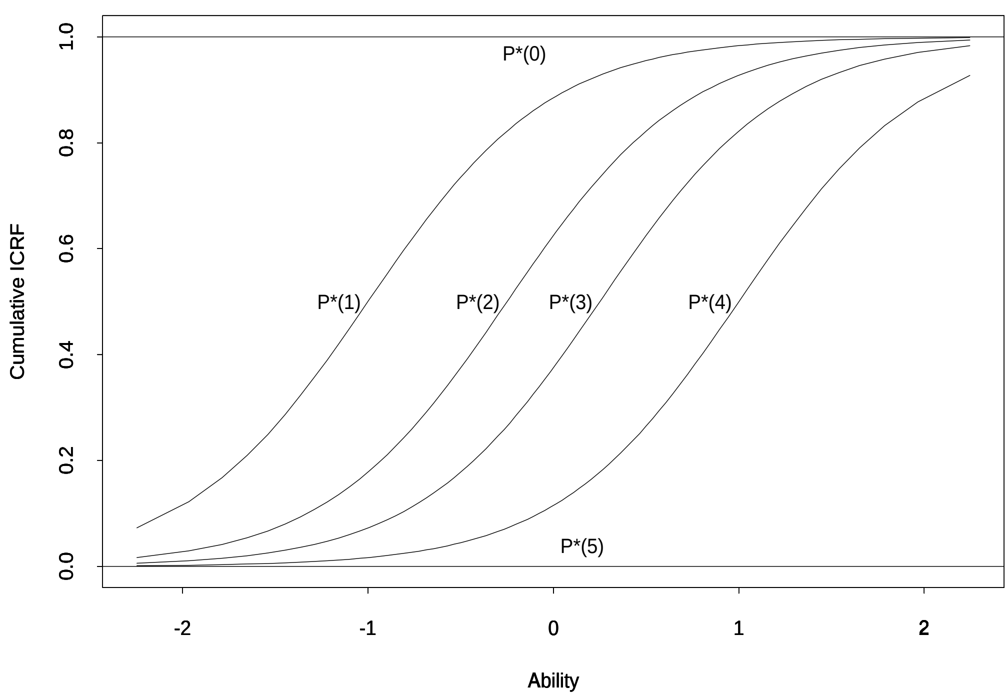
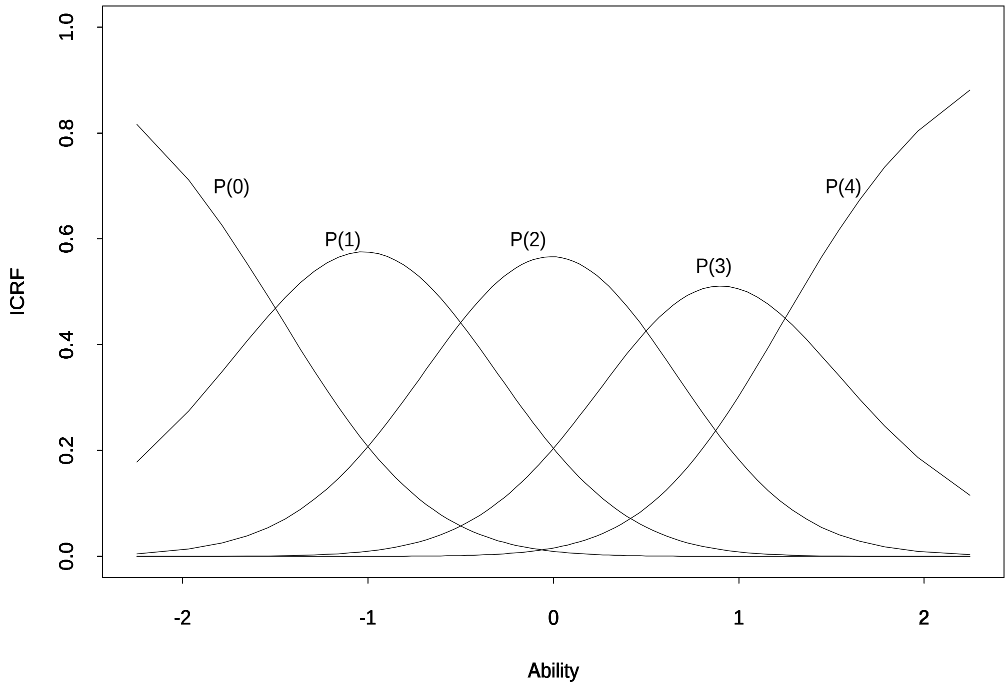


Figure 1.3: Generalized Partial Credit Model for Unidimensional Polytomous ICRFs



When the data are multidimensional, a generalization of the Generalized Partial Credit and Graded Response Models can be formulated in the same manner as the dichotomous multidimensional models. Thus, for the Graded Response Model, the probability a randomly sampled examinee with ability vector $\boldsymbol{\theta}$ obtains at least score q for $q = 1, \ldots, r_i$ is given by

$$P^{\star}_{i,q}(\boldsymbol{\theta}) = P(U_i \geq q | \boldsymbol{\Theta} = \boldsymbol{\theta}) = \frac{1}{1 + \exp[-1.7\mathbf{a}_i^T(\boldsymbol{\theta} - \mathbf{b}_{i,q})^T]}, \tag{1.17}$$

where the vector $\mathbf{a}_i = (a_{i,1}, \ldots, a_{i,d})^T$ consists of discrimination parameters for each dimension for item i and the vector $\mathbf{b}_{i,q} = (b_{i,q,1}, \ldots, b_{i,q,d})^T$ consists of difficulty parameters for each dimension for category q of item i.

Likewise, for the Generalized Partial Credit Model, the probability a randomly sampled examinee with ability vector $\boldsymbol{\theta}$ obtains score q for $q = 1, \ldots, r_i$ is given by

$$P_{i,q}(\boldsymbol{\theta}) = \frac{\exp\left[\sum_{v=1}^{q} 1.7\mathbf{a}_i^T(\boldsymbol{\theta} - \mathbf{b}_{i,v})\right]}{1 + \sum_{f=1}^{r_i} \exp\left[\sum_{v=1}^{f} 1.7\mathbf{a}_i^T(\boldsymbol{\theta} - \mathbf{b}_{i,v})\right]}, \tag{1.18}$$

where the vector $\mathbf{a}_i = (a_{i,1}, \ldots, a_{i,d})^T$ consists of discrimination parameters for each dimension for item i and the vector $\mathbf{b}_{i,q} = (b_{i,q,1}, \ldots, b_{i,q,d})^T$ consists of category parameters for each dimension for category q of item i.

1.3 Summary of Contents

One of the many areas of research in the field of Item Response Theory concerns the assessment of the dimensionality of a test. Developed by Stout (1987), the DIMTEST procedure is a nonparametric procedure that provides a hypothesis test of unidimensionality for a test data set. A new bias correction method for the DIMTEST procedure based on the nonparametric IRT bootstrap method (Kim, 1994; Gao, 1997) is introduced in Chapter 2. An estimate of each item's response function is obtained from the original data using the nonparametric method of kernel smoothing. Using a specified examinee ability distribution and the IRF estimates for each test item, a test data set is generated under the assumption of unidimensionality and another DIMTEST statistic is calculated from this generated data set. This process is repeated N times and the average of the N DIMTEST statistics is then subtracted from the value of the DIMTEST statistic calculated from the original data. Using results from both Stout (1987) and Douglas (1997), the new DIMTEST procedure is shown to have an asymptotically standard normal distribution under fairly general regularity conditions as both the number of items and the number of examinees tends to infinity. A simulation study in Chapter 2 shows this new version of the DIMTEST procedure has an average Type I error rate slightly below the nominal rate of $\alpha = 0.05$ and very high power to detect multidimensionality in a variety of realistic multidimensional models.

Chapters 3 and 4 contain extensions of the new bias correction method based on the non-parametric IRT bootstrap for the DIMTEST procedure to polytomous items and dichotomous items from a Computer Adaptive Test (CAT) respectively. For the Poly-DIMTEST procedure (Li, 1995), the bias correction method obtains an estimate of the ICRF for each item on the test using the non-parametric method of kernel smoothing. In a similar manner to DIMTEST, another test data set is generated under the null hypothesis of unidimensionality using a specified examinee ability distribution and the ICRF estimates for each test item. Another Poly-DIMTEST statistic is then calculated from the resampled unidimensional data set. This process is repeated N times and the average of the N Poly-DIMTEST statistics obtained from the N generated unidimensional data sets is subtracted from the value of the Poly-DIMTEST statistic calculated from the original data. A simulation study in Chapter 3 shows this new version of Poly-DIMTEST has an average Type I error rate slightly below the nominal rate of $\alpha = 0.05$ and very high power to detect multidimensionality in a variety of realistic multidimensional models.

In Chapter 4, the new bias correction method from Chapter 2 is applied to items from a Computer Adaptive Test (CAT). The form of the CAT administration considered in Chapter 4 is a multi-stage testlet design (Armstrong, Jones, Koppel & Pashley, 1999), currently under study by the Law School Admissions Council. Using this particular CAT administration and the new DIMTEST procedure from Chapter 2, the CAT-DIMTEST procedure is developed to test the null hypothesis of unidimensionality between the test's pretest items (items **not** used to estimate examinee ability on the test) and the test's operational items (items used to estimate examinee ability on the test). A simulation study in Chapter 4 shows the new CAT-DIMTEST procedure has an average Type I error rate slighly below the nominal rate of $\alpha = 0.05$ and very high power to detect multidimensionality when all or many of the pretest items are dimensionally different from the operational items. When the set of pretest items contains few items that are dimensionally different from the operational items, simulations show the power of the CAT-DIMTEST procedure decreases markedly.

This result underscores the importance of developing exploratory statistical procedures for detecting possible multidimensional items on a CAT.

Another area of research in the field of Item Response Theory is the estimation of item parameters and examinee abilities on a test for different unidimensional parametric IRF models. Many of the methods used to estimate these item and examinee parameters rely on the maximization of a specific likelihood function. The asymptotic theory of maximum likelihood estimates is well developed in the statistical literature when the number of parameters to be estimated is fixed as the sample size tends to infinity. However, when estimating both item parameters and examinee abilities, the number of parameters to be estimated increases with the sample size. Thus, standard theorems on consistency and asymptotic normality of maximum likelihood estimates do not apply to the estimation of both item parameters and examinee abilities.

The method most commonly used to estimate item parameters and examinee abilities on a test for the 2PL and 3PL unidimensional IRF models is the Marginal Maximum Likelihood Estimation (MMLE) method (Bock & Lieberman, 1970; Bock & Aitkin, 1981). The MMLE method is a two-stage procedure where item parameter estimates are obtained in the first stage and examinee ability estimates are obtained in the second stage. In order to prove the MMLE procedure produces consistent estimates of both item parameters and examinee abilities, the number of items and the number of examinees must both tend to infinity.

Using theorems from He & Shao (2000), Chapter 5 contains proofs of the consistency and asymptotic normality of the item parameters estimates obtained from the first stage of the MMLE procedure for the 1PL, 2PL and 3PL models as both the number of examinees and the number of items tends to infinity. The proofs in Chapter 5 depend upon fairly general regularity conditions on the model and on the growth of the number of items relative to the growth of the number of examinees.