Using Beaton Fit Indices to Assess Goodness-of-fit of IRT Models

By

Yutong Yin

MA, University of Connecticut, 2003

MA, University of Pittsburgh, 2005

Submitted to the Graduate Faculty of

School of Education in Partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

EDUCATION

This dissertation was presented

by

Yutong Yin

It was defended on

October 25, 2007

And approved by

Clement A. Stone, PhD, Associate Professor

Suzanne Lane, PhD, Associate Professor

Levent Kirisci, PhD, Associate Professor

Feifei Ye, PhD, Assistant Professor

Dissertation Director: Clement A. Stone, PhD, Associate Professor

Using Beaton Fit Indices to Assess Goodness-of-fit of IRT Models

Yutong Yin, PhD

University of Pittsburgh, 2007

The purpose of this study was to investigate the performance of Beaton's MR and MSR fit indices for assessing goodness-of-fit of IRT models. These statistics are based on a standardized residual calculated from an expected and observed response. The investigation was conducted using a Monte Carlo simulation study that varied conditions relevant to testing applications.

This research had three objectives: 1) To identify the sampling distribution of the fit statistics; 2) To assess the Type I error rates under different combinations of manipulated factors; and 3) To investigate the empirical power under different combinations of manipulated factors by introducing different types of model misfit.

The sampling distribution of Beaton's MR and MSR statistics belonged to the family of normal distribution. However, there was no basis for a theoretical normal distribution to test the hypothesis of model-data-fit. Therefore, Monte Carlo resampling methods were required to test the hypothesis of model-data-fit for Beaton's fit statistics.

Using Monte Carlo resampling methods for hypothesis testing, nominal Type I error rates were observed in this study regardless of test length, sample size, Monte Carlo resample size and number of replications. With regard to empirical power, higher power was observed for Beaton's MR statistic than MSR statistic under the condition that H_0 was false for the entire test. Under the condition that H_0 was false for a subset of test items, higher power for the misfitting item and more false rejections than expected for all the other items were obtained for Beaton's

MR statistic. In contrast, reasonable empirical power for the misfitting item and nominal Type I error rates for all the other items were observed for Beaton's MSR statistic. Based on the results of this study, Beaton's MSR fit statistics can be used to assess goodness-of-fit for both shorter (12 items) and longer test (36 items). The recommended sample size is 500 or more, and a Monte Carlo resample size of 100 should be adequate for hypothesis testing.

ACKNOWLEGEMENTS

First, I sincerely thank my advisor Dr. Clement Stone for giving me innumerable lessons and abundant guidance for the past four and half years, especially, for all his time and effort in this dissertation research. I feel so fortunate to have such a respectable advisor.

I would like to thank the rest of my committee members: Dr. Suzanne Lane, Dr. Feifei Ye, and Dr. Levent Kirisci. I appreciate their valuable suggestion and guidance on my dissertation. In particular, special thanks are given to Dr. Suzanne Lane, whose series of courses led me into the wonderful field of educational measurement.

I would also like to sincerely acknowledge Dr. Albert E. Beaton and Dr. Jie Li from Boston College for their generous help in the completion of this dissertation.

Finally, I would like to express my sincere gratitude to my parents, Fengqi Yin and Yunhua Ma for their endless love and support. Also, I would like to thank my husband, Yuqiang Huang and my daughter, Hannah Helena Huang. Thanks to my husband for the accompanying, you are the person behind my back whenever I feel tired or sad. Thanks to my little girl, and life with you is wonderful.

TABLE OF CONTENTS

LISTS OF TABLES

LISTS OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Statement of the Problem

Item response theory (IRT) has been widely used in educational and psychological measurement. An IRT model is a parametric model, which uses a mathematical formula to model the probability of a correct response with estimated ability and item parameters. The major merits of IRT over Classical Test Theory (CTT) are the properties of invariance of ability parameters across different tests and invariance of item parameters across different groups, which make IRT a test-free (estimate of examinee's ability does not depend on a particular test) and sample-free (estimate of item characteristics does not depend on particular group of examinees) measurement. So IRT can be used to solve a variety of measurement problems, such as selecting items, creating an item bank, equating tests from different test administrations, evaluating differential item functioning and implementing adaptive tests.

According to the number of response categories, IRT models have been classified as dichotomous and polytomous IRT models. For dichotomous IRT, there are 1PL, 2PL and 3PL models based on the item parameters in the model. For polytomous IRT, the Graded Response Model is most commonly used.

In order for an IRT model to be used in practice, a number of assumptions must be met. These assumptions include form of the IRT model, dimensionality, local independence and non-

speededness. In addition, a test to determine if the model fits the observed data is needed. Model-data-fit is important since the utility of an IRT model is dependent on the extent to which the model accurately reflects the observed data. Lack of fit between the model and the observed data will threaten the realization of IRT advantages. Lack of fit can be due to many reasons, including violation of the model assumptions and inadequacies in the estimation process.

1.2 Evaluating Model-data-fit

Since an IRT model is valid only when the model fits the data, the evaluation of model-data-fit is especially important. One way to evaluate the model-data-fit is to evaluate the degree to which the model predicts the observed item responses. IRT model-data-fit typically involves creating a two-way contingency table for each item, where the rows of a table correspond to discrete ability (θ) subgroups and the columns correspond to the possible score categories or response levels. An observed frequency distribution in the table is constructed by cross-classifying each examinee's ability level with his/her corresponding response to the item. An expected frequency distribution is obtained from the IRT model based probabilities of responses at each score level given the estimated item and ability parameters for each subgroup. Then, a test statistic or residual analysis can be used to compare the observed distribution with expected distribution to determine whether there is significant difference between the two distributions.

A number of traditional goodness-of-fit test statistics (Bock, 1972; Yen, 1991; Mckinley & Mills, 1985) have been proposed. These all compare an observed distribution with an expected distribution under a given model, and use a Pearson χ^2 statistic or likelihood-ratio G^2 statistic to test the hypothesis, where both statistics are assumed to follow a chi-squared distribution.

When using the above test statistics to assess model-data-fit, there are several disadvantages. First, sample size may affect the goodness-of-fit statistics and the hypothesized chi-square distribution for the test statistics. Sample size can affect the sensitivity of goodness-of-fit statistics. When sample size is small, serious misfit cannot be detected because of lack of statistical power. When sample size is large, any slight departure from the model would lead to rejection of the null hypothesis (Hambleton, 1989). Sample size will also affect the chi-square approximation of the distribution of goodness-of-fit statistics. In order for the fit statistics to approximate the chi-square distribution, there must be a large enough sample size. Sparse cells of the contingency table will also affect goodness-of-fit statistics. Sparseness happens in some cases even with a large sample size. Second, the performance of fit statistics is dependent on whether estimates of parameters are used. In IRT goodness-of-fit methods, each examinee is assigned to a specific ability subgroup based on ability estimates, and uncertainty in examinee ability estimates may result in misclassifications. The number of ability subgroups and the cut-points used to form subgroups are arbitrary, which may also affect the goodness-of-fit statistics (Reise, 1990). Lastly, all traditional goodness-of-fit methods assume a null chi-square distribution for fit statistics, and some simulation studies (Yen, 1981; Ansley & Bae, 1989 (as cited by Stone & Hansen)**;** Stone & Hansen, 2000) have suggested that these statistics are not always approximated well by a chi-square distribution. To improve the goodness-of-fit statistic, a number of alternative methods have been proposed.

Orlando and Thissen (2000) discussed a method that groups examinees based on examinee's observed total score rather than estimated ability. Orlando and Thissen's method compares the observed proportion with expected proportion in each score subgroup. The observed proportions are model-independent, which satisfies the assumption for asymptotic chi-

square distribution of Pearson χ^2 and likelihood-ratio G^2 statistics. They have also considered the influence of sparse cells with their goodness-of-fit method. However, Orlando and Thissen's method cannot be easily used to assess goodness-of-fit for polytomously scored item, and it still depends on the asymptotic chi-square distribution approximation for hypothesis testing.

Stone, Mislevy and Mazzeo (1994) described a method to account for the imprecision in ability estimation. In this method, the posterior ability distribution is used to classify an examinee into ability subgroups according to the probability that an examinee has ability equal to the subgroup. Thus, the uncertainty in ability estimation is considered directly. Then a pseudo-observed score distribution can be formed by summing the posterior probabilities for an item across all examinees. Goodness-of-fit statistics can be constructed based on the pseudo-observed score distribution and an expected score distribution. To test the null hypothesis, Stone (2000) found that the goodness-of-fit statistics were distributed as scaled chi-square distributions, and he discussed a rescaling method that could be used for hypothesis testing. However, the rescaling method still depends on a specified null chi-square distribution.

Residuals can also be used to evaluate the goodness-of-fit of IRT models. A residual is the difference between actual item performance for a subgroup of examinees and the subgroup's expected item performance. The traditional residual analysis of goodness-of-fit is a graphical procedure which visually compares a predicted and observed distribution. Beaton (2003) proposed an alternative method to assess goodness-of-fit which calculates standardized mean residuals and standardized mean squared residuals. Beaton fit indices still involve the comparison of an observed distribution with an expected distribution under a certain IRT model. Different from traditional methods, however, Beaton fit indices avoid the arbitration of using a specific number of ability subgroups and using cut-points to form ability subgroups. In Beaton fit

indices, the residuals for each examinee are based on the plausible ability values for each examinee. The plausible ability values are sampled from the Bayesian posterior ability for each examinee. Therefore, Beaton fit statistics account for the uncertainty in ability estimation. To test the null hypothesis, Beaton uses bootstrap resampling to simulate the empirical distribution and to determine if the IRT model is appropriate.

The present study evaluated the use of Beaton fit indices for assessing goodness-of-fit of IRT models. The goals of this study included:

(a) Investigate the sampling distribution of Beaton fit Statistics,

(b) Investigate the Type I error rates of Beaton fit statistics under different test conditions, and

(c) Investigate the empirical power of Beaton fit statistics under different test conditions.

The test conditions in this simulation study included different test lengths, sample sizes and Monte Carlo resample sizes.

1.3 Significance of the Study

IRT is increasingly used in educational and psychological measurement, and the decision to use a particular IRT model on a test dataset is crucial. Therefore, it is important to fully assess model-data-fit before applying any IRT model. Without evaluating model-data-fit, little is known about the appropriateness of a specific IRT model and the validity of various applications based on the specific IRT model may be threatened.

This study evaluated Beaton fit statistics using Monte Carlo simulations. The advantages of Monte Carlo simulations are that they can determine the sampling distribution of test statistics, and they can be used to manipulate factors and compare their effects on results.

Beaton's method is novel to assess the goodness-of-fit, but there is not much research evaluating its performance.

CHAPTER 2

REVIEW OF THE LITERATURE

2.1 Item Response Theory

Item response theory (IRT) uses ability and item parameters to model the probability of a correct response. It has been widely used in education and psychological testing. It has some advantages over Classical Test Theory (CTT), which was the dominant measurement theory prior to the 1980's. The basic idea of CTT is to decompose the observed score of an examinee into a true score and an error score. The advantages of CTT are that it relies on easily met assumptions, it employs relative simple mathematical procedures, and model parameter estimations are conceptually straightforward. However, CTT can be criticized for its limitations. First, the examinee score is test dependent. The true score is not an absolute characteristic of an examinee since it depends on the content of the test. For the same examinee, a simple test will result in a different score than a difficult test. This fact poses difficulty in comparing examinees who take different tests, or even different items within a test. Second, the item characteristics are group dependent. The item characteristics (e.g., item discrimination and item difficulty) depend on the sample of examinees that take a specific test. Dependence on the examinee group poses some major difficulties for test developers in applying CTT to some measurement situations (e.g., test equating and computerized adaptive testing). Finally, CTT is test oriented rather than

item oriented, and there exists no basis to predict how a given examinee will perform on a particular test item.

Item response theory (IRT) was originally developed to overcome the problems associated with CTT. Item response theory focuses on modeling the relationship between responses to an individual item and the underlying ability assumed to be measured by that item. Item response theory has three major advantages. First, ability estimation is independent of the test items being used, and therefore examinees can be compared even though they might not have taken the identical set of items. Second, item parameter estimation is group independent. Item statistics do not depend upon a particular group in a particular population of examinees and are assumed to be invariant across examinees. Finally, as the name implies, item response theory models responses at item level. Thus, an examinee's performance on test items can be predicted. The most important property of IRT is the invariance of item and examinee parameters. Because of this property, IRT provides a useful framework for solving a variety of measurement problems: building item banks, constructing new tests, equating scores from different test administrations, computer adaptive testing, etc.

IRT models are all based on specific assumptions about the data, and the validity of using IRT models depends on the degree to which these assumptions are met. The advantages of IRT over CTT are valid only if the assumptions of IRT can be satisfied. The assumptions of IRT (Hambleton, 1989; Hambleton & Swaminathan, 1991) are discussed as follows.

Form of the IRT Model. IRT is a model-based test theory. It uses mathematic functions to model the probability of a correct response. It assumes an examinee's performance can be predicted by one or more abilities. The correct response to an item has a monotonically increasing relation with the abilities. In other words, examinees with

higher abilities will correctly answer an item with a higher probability. The relation between observed responses and the abilities is modeled by item characteristic curves (ICCs). An ICC represents a nonlinear relation for the regression of item score on the abilities measured by the test. The shape of the ICC determines the mathematical function for the IRT model. An item response function (IRF) is a specific mathematical relationship between an examinee's performance, the examinee's abilities and test item parameters.

Dimensionality. The most commonly used IRT models assume that one single underlying ability or trait is sufficient to account for examinee performance. In real tests, however, this assumption cannot be strictly satisfied because several factors affect test performance. These factors include test motivation, test anxiety, speed of performance, test sophistication, and other cognitive skills. Although the above factors may be a component for an examinee to have a correct response in an assessment, it is sufficient for the unidimensionality assumption to be met adequately by a set of test data if one "dominant" component or factor influences test performance. This component or factor is referred to as the ability measured by the test.

Local independence. IRT assumes that item responses are conditionally independent, or an examinee's responses to different items in a test are statistically independent. LI specifies that only the examinee's ability and the characteristics of test items influence test performance. For this assumption to be true, an examinee's performance on one item must not affect, either for better or for worse, his or her responses to any other item on the test. For example, the content of an item must not provide clues to the answer to any

other test item. Local dependence can potentially arise among items that have a similar stem, items that have very similar content, and items that are presented sequentially.

Non-speededness. This assumption of IRT models means that tests are administered under non-speeded conditions. That is, any omitted test item is due to limited ability of examinees but not due to their failure to reach those items. When speed does affect test performance, the unidimensionality assumption is essentially violated since the trait measured by a test is not the only factor impacting test performance.

2.1.1 Dichotomous Item Response Theory

Dichotomous items have only two possible responses (e.g., true/false, agree/disagree, etc.). Three IRT models are commonly used for dichotomous items: one, two and three parameter logistic models.

The item response function (IRF) for the one parameter logistic (1PL) model is:

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

where $P_i(\theta)$ is the probability that an examinee with ability θ answers item i correctly,

a is the fixed item discrimination (slope) parameter,

b_i is the item difficulty (location) parameter for item i,

D is the scaling factor to make the logistic function as close as possible to the normal ogive function $D=1.702$.

As its name implies, the 1PL model uses a single item parameter, item difficulty (b_i). The item difficulty, b_i, is the value indicated by the midpoint between the lower and upper asymptotes of the IRF. The upper asymptote will always approach 1 as the ability level increases infinitely. For

1PL, the lower asymptote is always 0. Thus, the value of b_i is always indicated by the ability value at the point where the IRF is equal to 0.5.

Figure 2.1: the ICCs for 1PL Models

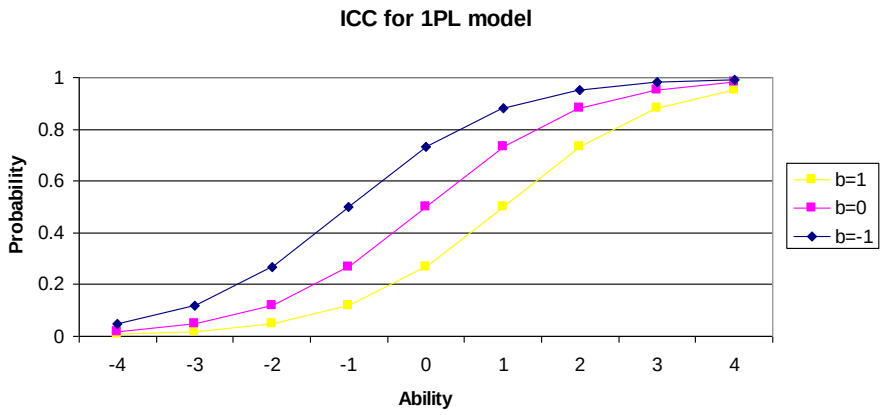


Figure 2.1 shows Item Characteristic curves (ICCs) for 1PL models ($b_1=1$, $b_2=0$ *and* $b_3=-1$). As shown in the figure, the probability of a correct response to an item increases monotonically by the ability level. In the middle of a curve, the probability increases sharply; at the extreme points, the probability changes slowly. All three ICCs have the same general shape, and they differ only in locations.

The 1PL model is appropriate for items that are equally related to the latent trait, since the model's distinguishing characteristic is that the discrimination power is the same for all items at different difficulty levels.

The 2PL model takes into account the variation in discrimination powers of test items. The item response function (IRF) for two parameter logistic (2PL) model is:

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

where a_i is the item discrimination (slope) parameter for item i, and $P_i(\theta)$, b_i and D have the same interpretations as in the 1PL model.

The discrimination parameter a_i is equal to the slope at the point of an ICC where the correct response probability is 0.5 and shows how well a test item can discriminate among examinees. Thus, items with steeper slope can separate examinees into different ability levels more easily than items with less steep slopes. For 2PL, the value of b_i is still indicated by the ability value at the point where the IRF is equal to 0.5. An example of ICCs for 2PL models is as follows.

Figure 2.2: ICCs for 2PL Models

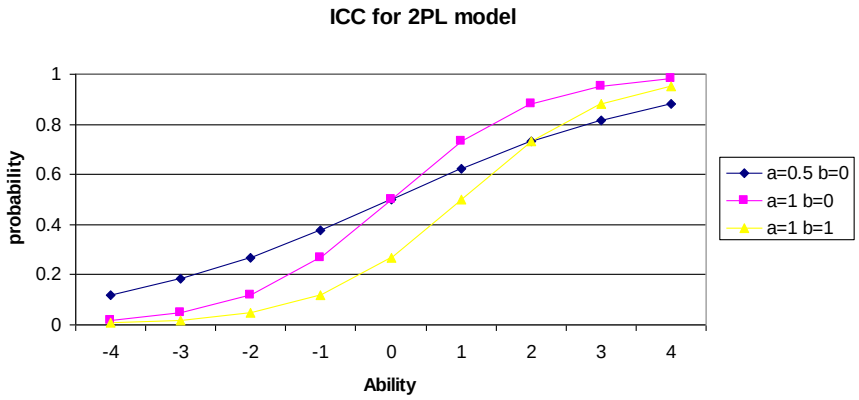


Figure 2.2 shows ICCs for 3 test items ($a_1=0.5$, $b_1=0$; $a_2=1$, $b_2=0$; $a_3=1$, $b_3=1$). Item 2 and item 3 have the same discrimination parameter ($a_2=a_3=1$) but different difficulty parameters ($b_2=0$, $b_3=1$). Thus, their ICCs have equal slope and hence never cross each other. In other words, for items 2 and 3, changes in ability level have an equal impact on the probabilities of correct response. Items 1 and 2 have the same difficulty ($b_1=b_2=0$) but different discrimination parameters ($a_1=0.5$, $a_2=1$). Item 2 is able to discriminate examinees more easily than item 1, as shown by the figure.

Within 1PL and 2PL models, the probability of a correct response approaches zero when the ability value approaches negative infinity ($-\infty$). The negative infinity situation happens when an examinee has almost no knowledge to correctly answer an item. However, an examinee may correctly respond by guessing, especially for multiple choice items. Under this circumstance, the probability of a correct response is greater than zero even for examinees with low ability. Faced with this problem, the 3PL model has been proposed to incorporate a guessing factor. The item response function (IRF) for the three parameter logistic (3PL) model is:

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

where c_i is the pseudo-guessing (lower asymptote) parameter for item i. $P_i(\theta)$, a_i, b_i and D have the same interpretations as in the 2PL model.

For the 3PL model, there is a nonzero pseudo-guessing parameter c_i. This "guessing" parameter characterizes the probability of examinees reaching a correct response simply by chance. c_i can be constructed as the percentage of correct responses from examinees of extremely low ability. c_i is a constant for a test item i.

Figure 2.3: ICC for a 3PL Model

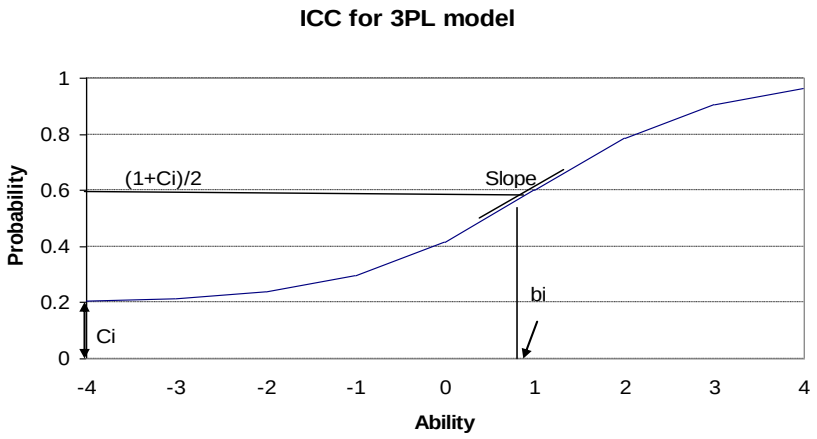


Figure 2.3 shows an example of ICC of a 3PL model. The pseudo-guess parameter (c_i) is the lower asymptote of an ICC. In other words, even examinees with extremely low ability still have some chance to answer an item correctly. For a 3PL model, the value of b_i is indicated by the ability value at the point where the probability of a correct response is $(1+c_i)/2$. The item discrimination parameter a_i is the slope at the point $\theta=b_i$.

2.1.2 Polytomous Item Response Theory

Rather than restrict the item responses to two categories, the polytomous IRT models allow the responses to be classified into more than two categories. For instance, a model, which contains more than two response categories, is necessary to measure examinee performance on a Likert scale item or to assign credit to a partially correct response. Another example is the multiple-choice item where the choice of different wrong answers reflects different ability levels. Thus, it is desirable to use a model that can assess information from all item options rather than use a model that only assumes an examinee either correctly responds or randomly selects an incorrect alternative. Several polytomous IRT models have been proposed. These include the Graded Response Model (Samejiam, 1969), the Partial Credit Model (Master, 1982) and the Nominal Response Model (Bock, 1972).

For example, the Graded Response Model (GRM) is appropriate for items having ordered response categories, where higher response categories indicate higher examinee ability. In the GRM, an examinee response falls in only one of the ordered categories for each item. A logistic function, called a boundary response function, is utilized by the GRM. The GRM is a direct extension of the two parameter logistic (2PL) model. The GRM uses the item response function

(IRF) of the 2PL model to define the boundary response function. Boundary responses of the GRM can be treated dichotomously. For instance, for an item with *m* categories, category *k (k<m)* and above can be treated as correct responses, and categories below *k* are incorrect responses. In this way, the boundary response function of the GRM has the same mathematic expression as the IRF of 2PL models. The boundary response function computes the probability that an examinee with ability θ will respond in category *k* and higher, and it is represented as:

$$P_{ik}^{*}(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_{ik})}},$$

where $P_{ik}^{*}(\theta)$ is the probability that an examinee with ability θ will respond in category *k* or higher for item *i*,

k=0, 1, ..., m is an response category for item *i*,

a_i is the item discrimination (slope) parameter for item *i*,

b_{ik} is the item difficulty (location) parameter for category *k* of item *i*,

D is the scaling factor to make the logistic function as close as possible to the normal ogive function *D=1.702*.

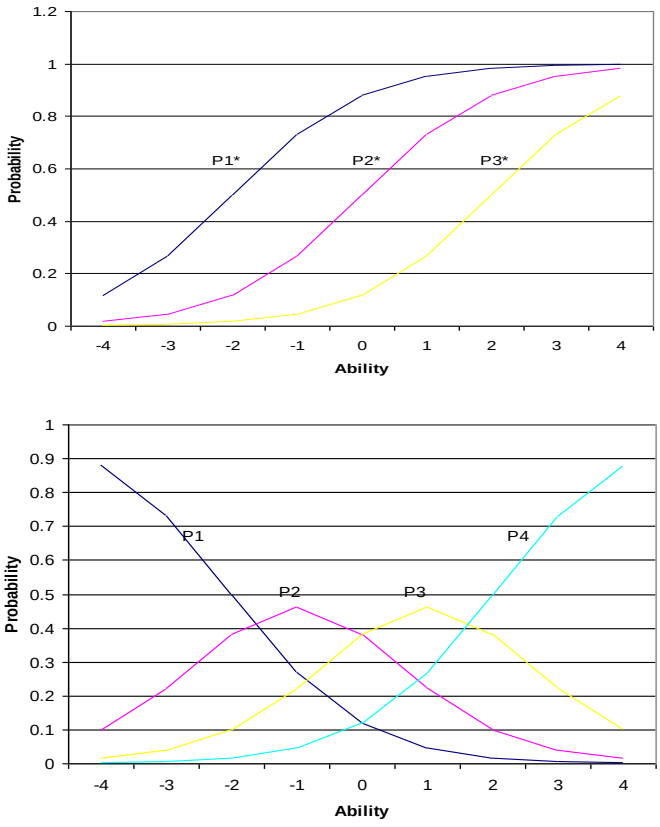
In the boundary response function, *m+1* is the number of ordered response categories for item *i*. The boundary response function is 1 for the lowest response category ($P_{i0}^{*}(\theta)=1$**)** and is 0 for the highest category ($P_{i(m+1)}^{*}(\theta)=0$**)**. The discrimination parameter a_i varies by item *i*, but a_i is the same for all response categories of an item. There are *m* item difficulty parameters (b_{ik}) for an item. The difficulty parameters are ordered as $b_{i(k-1)}<b_{ik}<b_{i(k+1)}$. The value of b_{ik} is the ability value (θ) at the point where the probability of responses in category *k* and higher is 0.5 (Thissen, 1991).

The category response function calculates the probability of responding in a particular category k, which is the difference between the boundary response function values for adjacent categories. For instance, the probability is P_{ik}^{*} for a response in category *k* and higher, and is $P_{i(k+1)}^{*}$ in category *k+1* and higher. Thus, the probability of a response in category *k* is:

$$P_{ik}(\theta) = P_{ik}^{*}(\theta) - P_{i(k+1)}^{*}(\theta),$$

The boundary response functions $P_{ik}^{*}(\theta)$ on θ are plotted as Operating Response Curves (ORCs) in figure 2.4, and the category response functions $P_{ik}(\theta)$ on θ are drawn as Category Response Curves (CRCs). Figure 2.4 presents an example of ORCs and its corresponding CRCs with three response categories (*a=1.0, b_1=-2.0, b_2=0, b_3=2.0*) for a polytomous item.

Figure 2.4:
Operating Response Curves and Category Response Curves for an Item with 3 Categories



As shown in Figure 2.4, ORCs have the same S shape but are different in location (difficulty). The value of the difficulty parameter (b_{ik}) is the ability value at the point where the probability of responding in category k and above is 0.5. In CRCs, b_{ik} is referred to as the ability value at the peak point for intermediate response categories. The item discrimination parameter (a_i) is the slope of any ORC at the point where b_{ik} is located. In general, the higher the discrimination parameter (a_i), the steeper the ORC and associated CRC will be narrower and higher. This indicates that the response categories differentiate among ability levels fairly well. However, CRCs of all response categories lack a consistent pattern. This situation makes the modeling of the CRC and parameter estimation complicated.

2.2 Model-data-fit

Assessing model-data-fit is important in item response theory (IRT) since the power of an IRT model is realized only when an adequate fit between the model and the dataset of interest is clearly established. Lack of fit may occur due to many reasons. First, the model assumptions may not be met by the data. For example, in IRT, the shape of the function relating performance to ability is assumed to be fixed. If the assumed function shape does not represent the observed data, misfit occurs. For instance, when guessing is a factor in the observed data, the three-parameter should be used to model the data. But, if a one-parameter model is used, misfit occurs. Second, there may be inadequacies in the estimation process. This can occur with small sample sizes, poor estimation algorithms, or a variety of other problems such as nonmontonicity of item-trait relations or poor item construction (Mckinley & Mills, 1985).

Misfit can affect the evaluation of student achievements in a variety of ways. For example, Hambleton and Cook (as cited by Hambleton, 1989) considered the effect of using an

incorrect model to obtain ability estimation for ranking examinees. Other researchers have studied the effect of parameter estimation errors on equating and adaptive testing.

There are two standard methods for assessing model-data-fit: statistical chi-square tests (Bock, 1972; Yen, 1991; Mckinley & Mills, 1985; Stone, Mislevy & Mazzeo, 1994; Orlando & Thissen, 2000) and residual analysis (Hambleton, 1989). Both statistical chi-square tests and residual analysis are obtained by comparing the actual item performance against the predicted item performance given the assumed IRT model.

Goodness-of-fit of an IRT model is best illustrated by using two-way contingency tables, where the rows of the table are defined to be ability subgroups (θ) and columns defined to be score response categories. Almost all goodness-of-fit methods are composed of six steps. First, estimate item and ability parameters by fitting an IRT model to the data. Second, create ability subgroups by dividing the continuous ability distribution into a small set of discrete intervals. Third, construct an observed score response distribution by cross-classifying examinees to one cell of a two-way table using their ability estimates and score responses. Fourth, construct an expected score response distribution based on the probability of each response by using item parameter estimation and subgroup ability. Fifth, compute a Chi-square test statistic or residual by comparing the observed and expected distributions. Finally, test the null hypothesis that the model fits the model (Stone, 2000C).

An item fit table can be displayed as a two-way contingency table, which includes both the observed and expected score distribution (Stone & Hansen, 2000). For example, Table 2.1 presents a fit table for an item with 3 score levels. In the table, O_{kj} and E_{kj} are the observed and expected frequencies, respectively, for individuals with ability level k and response score levels $j = 0, 1$ and 2. Note that the information in the table basically involves a discrete ability level

comparison of the empirical item responses with the modeled item responses. The item fit table can be used for both dichotomous and polytomous items. For dichotomous items, the responses only include 0 and 1.

Table 2.1
Cross-Classification Table of Ability Level and Score Response for an Item with Three Score Levels

θ Group	Score Response			
	0	1	2	total
1	O_{10} (E_{10})	O_{11} (E_{11})	O_{12} (E_{12})	$O_{1.}$ ($E_{1..}$)
2	O_{20} (E_{20})	O_{21} (E_{21})	O_{22} (E_{22})	$O_{2.}$ ($E_{2..}$)
3	O_{30} (E_{30})	O_{31} (E_{31})	O_{32} (E_{32})	$O_{3.}$ ($E_{3..}$)
.				
.				
.				
K	O_{k0} (E_{k0})	O_{k1} (E_{k1})	O_{k2} (E_{k2})	$O_{k.}$ ($E_{k..}$)
	$O_{.0}$ ($E_{.0}$)	$O_{.1}$ ($E_{.1}$)	$O_{.2}$ ($E_{.2}$)	

2.2.1 Goodness-of-fit Statistics

The goodness-of-fit methods for IRT models are typically based on a Pearson χ^2 statistic or a likelihood-ratio G^2 statistic.

The Pearson (1900) χ^2 statistic is the foundation to assess goodness-of-fit for IRT models and is used to test how far observed frequencies deviate from expected frequencies. The Pearson χ^2 is defined as:

$$\chi^2 = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{(O_{kj} - E_{kj})^2}{E_{kj}},$$

where k is the row category,

K is the number of row categories in a table,

j is the column category,

J is the number of column categories in a table,

O_{kj} is the observed frequency for cell kj, and

E_{kj} is the expected frequency for cell kj.

For the Pearson χ^2 statistic to be valid, some assumptions must be met. First, the observations must be independent. Each observed response can contribute to only one cell. Second, the sample size of observed data should be large. A large sample can guarantee that Pearson χ^2 statistic has an asymptotic chi-square distribution. Furthermore, the denominator in the χ^2 is the expected frequency, which may be decreased as the sample size decreases. If the expected frequency in some cells is too small, the value of χ^2 would be overestimated and would result in rejecting the null hypothesis. When all assumptions are satisfied and E_i is not dependent on estimated parameters, the Pearson χ^2 statistic is asymptotically chi-square distributed with degrees of freedom as *(K-1)*(J-1).* If E_i depends on estimated parameters, the correct degrees of freedom are *(K-1)*(J-1)-p*, where p is the number of estimated parameters (Fisher, 1924).

A likelihood-ratio G^2 statistic for goodness-of-fit was proposed by Neyman and Pearson (1928). It involves the ratios between the observed and expected frequencies. The likelihood-ratio G^2 statistic is defined as:

$$G^2 = 2\sum_{k=1}^{K}\sum_{j=1}^{J} O_{kj} \log\left[\frac{O_{kj}}{E_{kj}}\right],$$

where k, K, j, J, O_{kj} and E_{kj} have the same interpretations as in Pearson's χ^2 statistic.

The assumptions associated with the likelihood-ratio G^2 statistic are the same as those with Pearson χ^2 statistic. When all the assumptions are satisfied, G^2 is also asymptotically chi-

square distributed with degrees of freedom as *(K-1)*(J-1) or (K-1)*(J-1)-p*, depending on whether the E_i is estimated or not.

2.2.2 Traditional IRT Goodness-of-fit Statistics

To assess the goodness-of-fit of IRT models, variations of the Pearson χ^2 statistic and likelihood-ratio G^2 statistic include Bock's χ^2, Yen's *Q1* and Mckinley & Mills's (1985) G^2 statistic.

Bock's χ^2 statistic (1972)

Bock proposed a Pearson chi-square statistic to examine the suitability of an IRT model. Bock's fit statistic has been known as Bock's χ^2, and is defined as follows:

$$\chi^2 = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{N_k (O_{kj} - E_{kj})^2}{E_{kj}(1 - E_{kj})},$$

where k is the number of an ability subgroup,

K is the total number of ability subgroups,

j is the response score level,

J is the maximum response score level,

N_k is the number of examinees within ability subgroup k,

O_{kj} is the observed proportion of responses j within ability subgroup k,

E_{kj} is the expected proportion of responses j within ability subgroup k.

For an item, the ability θ scale is divided into K intervals, where roughly equal numbers of examinees can be placed into each interval according to the rank order of ability levels. The observed response proportion (O) is based on the response in each ability interval. The expected

correct response proportion (*E*) is estimated by using item parameter estimates and the median of ability estimates in each ability interval.

Bock's χ^2 is assumed to be chi-square distributed with degrees of freedom as *K*(J-1)-m*, where *m* is the number of estimated item parameters. The degrees of freedom for Bock's χ^2 are different from the goodness-of-fit statistics discussed earlier. Since IRT reflect a latent trait model, the ability θ subgroups *(k)* are independent, and there is no restriction that $\sum_k E_{kj} = 1$ across all subgroups *(k=1* to *K)* for a specific score level (Yen, 1981; Stone and Hansen, 2000). Therefore the degrees of freedom for the number of ability subgroup (*K*) do not need to be adjusted by 1. For dichotomous items with *j* equal to 2, the degrees of freedom are reduced to *K-m*.

Yen's Q1 Statistic(1981)

Bock's χ^2 statistic was modified by Yen (1981). Yen's statistic is referred to as *Q1*, and defined as follows.

$$Q_1 = \sum_{k=1}^{10} \sum_{j=1}^{J} \frac{N_k (O_{kj} - E_{kj})^2}{E_{kj}(1 - E_{kj})},$$

where *k, j, J* , N_k, O_{kj} and E_{kj} have the same interpretations as Bock's χ^2 statistic.

The construction of Q_1 is very similar to Bock's χ^2 with two exceptions. First, examinees are placed into 10 intervals according to the rank order of ability levels. But, the number of intervals is not specified in Bock's χ^2 statistic. Second, E_{kj} is the mean of the probabilities of response *j* for examinees within ability subgroup *k.* Q_1 is approximately distributed as a chi-square distribution with degrees of freedom as *10*(J-1)-m*, where *m* is the number of estimated item parameters. For dichotomous item, the degrees of freedom are *10-m*.

Likelihood-ratio G^2 (Mckinley & Mills, 1985)

Mckinley & Mills (1985) constructed a likelihood-ratio G^2 statistic based on computations similar to those used for Q_1. G^2 is defined as follows.

$$G^2 = 2\sum_{k=1}^{10}\sum_{j=1}^{J} O_{kj} \ln\left(\frac{O_{kj}}{E_{kj}}\right),$$

In the equation, *k, j, J,* O_{kj} and E_{kj} have the same interpretations as Bock's statistic. The likelihood-ratio G^2 is approximately distributed as a chi-square distribution with degrees of freedom as *10*(J-1)-m*, where *m* is the number of estimated item parameters. For dichotomous item, the degrees of freedom are *10-m*.

2.2.3 Limitations of Traditional IRT Goodness-of-Fit statistics

Although these traditional methods are useful for detecting various types of model misfit, there are several limitations associated with these methods.

2.2.3.1 Effect of Sample Size and Sparseness on Goodness-of-fit Statistics.

Since the traditional IRT goodness-of-fit statistics are all based on Pearson χ^2 and likelihood-ratio G^2 statistics, the limitations related to Pearson χ^2 and likelihood-ratio G^2 statistics will still affect the traditional IRT goodness-of-fit statistics.

One well known limitation of goodness-of-fit statistics is related to sample size, since a large sample size is a requirement for Pearson χ^2 and likelihood-ratio G^2 statistics to be asymptotic chi-square distributed. Therefore, these traditional IRT goodness-of-fit methods require large sample sizes. At the same time, however, Pearson χ^2 and likelihood-ratio G^2

statistics are sensitive to examinee sample size, which is one criticism of the two statistics. When sample size is small, serious departures from the null hypothesis may still not be detected. When sample size is very large, small and unimportant departures from the null hypothesis are almost certain to be detected.

Sparseness is a further limitation related to sample size. The term "sparse" refers to a circumstance that there are one or more cells with small frequencies. Sparseness occurs when the sample size is small, or when the sample size is large but there are also a large number of categories. Sparse data implies small expected cell frequencies *(E)*. When the data are sparse, Pearson χ^2 and likelihood-ratio G^2 statistics may not validly assess goodness-of-fit. There are a variety of different opinions on how small the expected frequency can be without invalidating the chi-square approximation. Fisher (1941) recommended that no expected frequency for a category can be less than 5. Cramer (1946) pointed out that each expected frequency needs to be at least 10. Kendall (1952) stated that the approximation may confidently be applied when each expected frequency is no less than 20. A common agreement is that the expected frequency for each category is 5 or more for a small number of categories (e.g., four), or the expected frequency for 80% of the categories is 5 or more for a large number of categories, but no frequency is zero.

Sparseness in the tables will affect the null hypothesis distribution, type I error and empirical power of Pearson χ^2 and likelihood-ratio G^2 statistics.

Null distribution: When the null hypothesis is true and there is no sparseness, the distributions of Pearson χ^2 and likelihood-ratio G^2 are both approximated by a chi-square distribution. Under the condition of data sparseness, the approximate chi-square distribution does not agree well with the exact distribution of Pearson χ^2 and likelihood-ratio G^2 statistics.

The Pearson χ^2 is more closely distributed as a chi-square distribution than the likelihood-ratio G^2 statistic. The distribution of the likelihood-ratio G^2 statistic is usually approximated poorly by a chi-square distribution when expected cell frequencies are less than 5.

Type I error rate: The Pearson χ^2 is a better statistic than the likelihood-ratio G^2 statistic in terms of having type I error rates that are closer to nominal levels based on the asymptotic chi-square approximation. Type I error rates for the Pearson χ^2 are close to the nominal level for a wide range of sample size while type I error rates for the likelihood-ratio G^2 statistic are too high even with moderate cell expectations. Larntz (1978) showed that the Pearson χ^2 statistic was more robust to small expected frequencies than the likelihood-ratio G^2 statistic. Larntz reported that the Pearson χ^2 achieved the desired rejection rate under the null hypothesis when all expected cell frequencies were greater than 1.0. The likelihood-ratio G^2 statistic, however, was much more sensitive to small cell expectations. Larntz showed that when the null hypothesis was true and when expected cell frequencies were smaller than 0.5, type I error rates for the likelihood-ratio G^2 statistic were much less than the expected nominal rates (α). When cell expected frequencies were between 1.5 and 4.0, however, the likelihood-ratio G^2 statistic rejected the null hypothesis too often (Agresti, 1990). Davier's (1997) simulation compared the type I error rates for both the Pearson χ^2 and the likelihood-ratio G^2 statistics under different conditions. When the data were sparse, the likelihood-ratio G^2 exhibited extremely high type I error rates. When the data were not sparse, the likelihood-ratio G^2 statistic showed similar type I error rates as the Pearson χ^2 statistic. The Pearson χ^2 statistic showed consistent type I error rates when the data were sparse or not sparse.

Empirical power: The Pearson χ^2 is a better statistic than the likelihood-ratio G^2 statistic in terms of having higher power to detect misfit based on the asymptotic chi-square

approximation. Lartza (1978) also investigated the empirical power of the Pearson χ^2 statistic and the likelihood-ratio G^2 statistic. He found the empirical power of the Pearson χ^2 statistic was always higher than the likelihood-ratio G^2 statistic even when the minimum cell expected frequencies were between 1.5 and 4.0. Davier (1997)'s simulation also compared Empirical power for both the Pearson χ^2 statistic and the likelihood-ratio G^2 statistic under different condition, When the data were sparse, the likelihood-ratio G^2 statistic exhibited Low empirical power than expected, when the data were not sparse, the likelihood-ratio G^2 statistic showed similar empirical power as the Pearson χ^2 statistic. The Pearson χ^2 statistic achieved enough power when the data were sparse or not sparse.

Some remedy methods have been proposed to solve the problem of sparseness. These include collapsing categories and adding a small constant to each cell.

Collapse categories. When frequencies are too small to permit a chi-squared approximation, researchers often combine categories until the combined expected frequency become large enough for the chi-square approximation to apply. But collapsing categories may not be a good idea for several reasons. First, different methods of collapsing categories will yield different results for evaluating fit statistics. Second, collapsing categories may make some categories in the table dependent on each other, thus violating the independent assumption of chi-square tests. Finally, collapsing too many categories will reduce statistical power of a test, since we need to subtract one degree of freedom when collapsing a category. Collapsing categories is suitable in situations when there is a natural way to combine categories and little information is lost when defining variables more crudely (Agresti, 1990).

Adding a small constant to each cell. Sparse table usually contain empty cells, cells with zero frequency. When empty cells are present, adding a small constant to each cell is

sometimes recommended. This remedy method may be helpful if only a few cells have low expected frequencies, but it is not useful when there are many sparseness cells in the table. Adding a constant, like 0.5, can significantly increase the sample size for a large table, and it smoothes the data too much. In even moderately sparse tables, adding a small constant often produces a large conservative effect on the outcome of goodness-of-fit (Agrest, 1990). It is difficult to decide which specific constant to add. However, the sum of added constants should be a very small percentage of the total sample size of observed data. Agrest recommended, if there is a problem with computation, a very small constant (.000001) should be added to avoid over-smoothing of the data.

2.2.3.2 Limitations Related to Assessment of Goodness-of-fit of IRT Models

IRT models are latent models, where true ability θ is unknown and has to be estimated. Reise(1990) pointed out the problem with estimated ability. That is, since the number of ability subgroups and the cut-points used to form ability subgroups are arbitrary, different subgroup partitions may generate different goodness-of-fit test results. When choosing cut-points to form ability intervals, the intervals should be wide enough so that the number of examinees in each interval is not too small. Small intervals can lead to unstable statistics. But the interval for each ability subgroup should not be too wide to maintain the similarity among examinees within that interval. Mckinley and Miller's G^2 and Yen's *Q1* used 10 ability subgroups to assess goodness-of-fit of an IRT model. Although Yen (1981) showed that the use of 10 ability subgroups was ideal in most test situations, it is still arbitrary and the value of *Q1* statistic may be influenced by the number of ability subgroups. Another problem is test length. Although an examinee's estimated ability is not dependent on the particular sample of test items, the precision of the

examinee's estimated ability is highly dependent on the length of the test. Longer tests are always associated with more precise of ability estimates than shorter tests. For shorter tests, due to the imprecision of ability estimation, classification error will probably occur when an examinee is assigned to a wrong subgroup.

Evaluating the null distribution. For the traditional goodness-of-fit statistics, when the null hypothesis is true, the test statistics are assumed to follow a chi-square distribution. A number of research studies have been conducted to investigate the null distribution of the traditional goodness-of-fit statistics. Yen (1981) investigated the distribution of *Q1*, and found that the mean of *Q1* was always greater than the degrees of freedom (*10-m*) for a dichotomous item. Since the mean of chi-square distribution should be equal to the degrees of freedom, Yen concluded that *Q1* was approximately distributed as a chi-square distribution with some distortion.

Ansley & Bae (as cited in Stone & Hansen, 2000) also carried out a simulation study to investigate the sampling distribution of Yen's *Q1* statistic for the 3PL IRT model. The simulation study included conditions of different test length (30 and 60 items) and examinee sample size (1000 and 2000). They found that the Q1 statistic was distributed as a non-central chi-square distribution. The non-centrality parameter varied with sample size and test length. For a given test length, the non-centrality parameter increased along with increased sample size, which indicated more severe deviation from a chi-squared distribution. For a given sample size, the non-centrality parameter decreased along with increased test length, which indicated a closer approximation to a chi-square distribution.

Stone and Hansen (2000) investigated the null sampling distribution of the Pearson and the likelihood ratio chi-square statistics for item response data under a 5-category GRM. The

empirical sampling distributions were evaluated under conditions that varied test length (8, 16 and 32) and sample size (1000 and 2000). They compared the means and variances of the sampling distributions, Q-Q plots and type I error rates under different combination of conditions. Result indicated that, for a test length of 32 items, the sampling distributions of the statistics approximated the null chi-square distribution fairly well. For tests that consisted of 8 and 16 items, results showed more departures in the test statistic from the null chi-square distribution.

All these studies indicate that there is some uncertainty about the null distribution of traditional goodness-of-fit statistics.

Examining Type I error. Type I error is false rejection rate of a correct null hypothesis. To evaluate Type I error rate, researchers (Mckinley and Mills, 1985; Orlando and Thissen, 2000 and 2003; Stone and Hansen, 2000b) investigated goodness-of-fit of IRT models under the different conditions. The work by Mckinley and Mills compared traditional goodness-of-fit methods (Mckinley and Mills's G^2, Bock's χ^2 and Yen's *Q1*) for a test of 75 items and three sample sizes 500, 1000 and 2000. Type I error rates were evaluated by simulating and calibrating data using the same or higher order IRT model (e.g. if simulate data using 2PL model, then the data was calibrating using 2P or 3P model). Their results showed that the sample sizes of 500 and 1000 yielded fewer false rejections than 2000. All of the three statistics showed similar results, the Mckinley and Mills's G^2 yielded fewer false rejections than Bock's χ^2 and Yen's *Q1*.

The work by Orlando and Thissen (2000) studied the performance of Yen's *Q1* and the likelihood-ratio G^2 statistic. They compared type I error rates for tests of 10, 40 and 80 items with fixed sample size of 1000. The result showed that all of the tests had quite high type I error rates. For an expected nominal α of .05, empirical α was around 0.95 for the test of 10 items,

between 0.10 and 0.29 for the test of 40 items, and somewhat lower but still inflated for the test of 80 items. Therefore, we can see that: type I error rates in traditional methods were inflated for short tests. The performance of these methods improved as test length increased and would approach the nominal rate if the test was long enough.

Stone and Zhang (2003) examined Bock's χ^2 statistic under different conditions of test length (10, 20 and 40) and sample size (500, 1000 and 2000). They found the type I error rates increased with increased sample size. For a 10 items test, the type I error rates increased from 0.84 to 1.00 when sample size increased from 500 to 2000. The type I error rates were so high that all items were identified as misfit.

Examining Empirical Power. Empirical power is the correct rejection rates of a false null hypothesis. Empirical power is evaluated for IRT fit statistics when models generate data by utilizing different number of item parameters than the number used in calibrating the data. From Mckinley and Mills's (1985) studies, Empirical power was evaluated by using higher order IRT model to simulate data and calibrating using the lower order IRT model. (e.g. if simulate data using 3PL model, then the data was calibrating using 1P and 2P model). The result showed that all these three statistics achieved a similar power in detecting misfitting items, Bock's χ^2 yielded fewer false accept than Mckinley and Mills's G^2 and Yen's *Q1*. The results also illustrated that empirical power increased as sample size increased from 500 to 2000. Orlando and Thissen (2000) investigated the empirical power of traditional methods on short tests. Their results showed that empirical power for Yen's *Q1* and likelihood-ratio G^2 Statistic were not useful since the two statistics both have highly inflated type I error rates.

2.2.4 Alternative Goodness-of-fit Statistics

Because the limitations of the traditional statistic methods for goodness-of-fit of IRT models, a number of alternative methods have been proposed. Three of these new methods will be discussed.

2.2.4.1 Fit Statistics Conditioning on Total Score

Instead of using ability estimates to divide examinees into subgroups, Orlando & Thissen (2000) partitioned examinees into subgroups based on the observed total test score. They proposed two new fit statistics (*S*-χ^2 and *S*-G^2) based on traditional χ^2 and G^2 statistics.

The *S*-χ^2 statistic is defined as:

$$S-\chi_i^2=\sum_{k=1}^{n-1}N_k\frac{(O_{ik}-E_{ik})^2}{E_{ik}(1-E_{ik})},$$

where N_k is the number of examinees at score subgroup *k*,

O_{ik} is the observed proportion of correct response to item *i* by subgroup *k*, and

E_{ik} is the expected proportion of correct response to item *i* by subgroup *k*,

The *S*-G^2 statistic is defined as:

$$S-G_i^2=2\sum_{k=1}^{n-1}N_k\left[O_{ik}\ln\left(\frac{O_{ik}}{E_{ik}}\right)+(1-O_{ik})\ln\left(\frac{1-O_{ik}}{1-E_{ik}}\right)\right],$$

where the notations have the same interpretations as those in *S*-χ^2. Same as traditional χ^2 and G^2, *S*-χ^2 and *S*-G^2 are also approximately chi-squared distributed.

The observed proportion of correct responses to an item by a subgroup is the percentage of examinees that correctly responded to the item in that subgroup. Because of subgroup division based on the observed total test scores, the traditional way of calculating expected proportions is

not applicable. To solve this problem, they proposed a novel method to compute an expected proportion, that involves the joint likelihood of correct response to item *i* and a total test score *k* divided by the marginal likelihood of a total test score *k*. The equation is defined as follows:

$$E_{ik} = \frac{\int P_i(\theta) S_{k-1}^{*i} \phi(\theta) d\theta}{\int S_k \phi(\theta) d\theta},$$

where *k* is a total test score,

S_k is the likelihood of a total test score *k*,

$P_i(\theta)$ is the probability of a correct response to item *i*,

S_{k-1}^{*i} is the likelihood of a total test score *k-1* without item *i*, and

$\Phi(\theta)$ is the prior normal probability density at θ.

To calculate S_k and S^{*i}_{k-1}, a recursive algorithm was used. It starts with evaluating the likelihood with only one item, which are the probabilities of incorrect ($S^*_0=1-P_1$) and correct ($S^*_1=P_1$) response to item 1. Then, the likelihood is calculated with an item added each time until the last item is included. For instance, when item *i* is considered, the likelihood for score *k* is calculated as:

$$S_k = P_i S_{k-1}^* + (1 - P_i) S_k^*,$$

where S^*_{k-1} is the likelihood for total score *k-1* without item *i*,

S^*_k is the likelihood for total score *k* without item *i*, and

P_i is the probability of a correct response to item *i*.

The Orlando and Thissen's method seems promising since the observed proportions are solely a function of observed total test scores and examinees do not need to be sub-grouped in an arbitrary and model-dependent manner. The disadvantage is that the method requires comparable total test scores from examinees. Thus, it is not suitable for some testing applications, such as

computerized adaptive tests in which examinees receive different tests. This procedure has also yet to be extended to applications in valuing polytomous items.

Orlando and Thissen evaluated the performance of S-χ^2 and S-G^2. The type I error rates for S-χ^2 were close to the nominal rejection rate 0.05 and not affected by the test length. The type I error rates for S-G^2 were higher than for S- χ^2. Furthermore, S-G^2 performed poorly for long tests. The reason hypothesized by these researchers was that there usually are a larger number of score subgroups in longer tests and the number of subgroups with few member increases. Thus, the number of subgroups with small expected correct response proportions increases, which affects S-G^2 performance. As shown by empirical power, S-χ^2 is promising for detecting item misfit. S-G^2 is not very useful because of the inflated type I error rate.

2.2.4.2 Fit Statistic Based on Posterior Expectations

Stone, Mislevy and Mazzeo (1994) provided a method based on posterior expectations to account for inaccurate ability estimation. This is particular relevant in testing application where ability estimates are imprecise (e.g. short tests). Similar to traditional methods, Stone's method still calculates Pearson χ^2 and likelihood ratio G^2 statistics. The difference in Stone's method from traditional methods involves the way in which the item fit tables are constructed. In traditional methods, point estimates of ability are used to construct the item fit table. In Stone's method, an examinee's posterior ability distribution is used to construct the item fit table. The idea is to use conditional probability of unknown quantity based on the observed dataset. In this method, Bayes's theorem is used to predict the posterior ability distribution. Bayes's theorem is expressed as:

$$P(A/B) = P(B/A)P(A)/P(B)$$

In the IRT context, the conditional probability relates the unknown quantity of ability θ for an examinee with the examinee's item responses. Thus, the posterior probabilities of ability can be expressed as:

$$P(\theta / x) = P(x / \theta)P(\theta) / P(x),$$

where $P(\theta/x)$ is the posterior probability distribution of θ,

$P(x/\theta)$ is the conditional probability of response pattern x given θ,

$P(\theta)$ is the prior ability distribution, and

$P(x)$ is the marginal probability of response pattern x for an examinee with unknown θ randomly sampled from a population with a given distribution.

The posterior distribution combines information from prior ability distribution (assume the prior ability distribution is $N(0,1)$) and the likelihood function $P(x/\theta)$. The posterior distribution is the ratio of the joint distribution of x and θ and the marginal distribution of x and θ. The marginal distribution was used to standardize the likelihood function so that the area under the function is equal to 1.

Because the continuous θ scale cannot be evaluated analytically, they used a set of discrete quadratic points to approximate the continuous θ scale. The posterior probabilities at each score level j given θ level k are estimated by:

$$P_{jk} = \sum_{n=1}^{N} x_{jk} P(x_n / X_k) A(X_k) / P(x_n),$$

where P_{jk} is the posterior probability of an item for score level j and θ level k,

n is the nth examinees in the sample,

N is the number of total examinees,

x_{jk} is an indicator variable, which is *1* if the observed response of examinee n is equal to j and is *0* otherwise,

$P(x_n/X_k)$ is the conditional probability of n^{th} examinee's response pattern (x) at the k^{th} quadratic point (X) of the θ distribution,

$A(X_k)$ is the weight at the quadratic point X_k, and

$P(x_n)$ is the marginal probability of observed response pattern x for n^{th} examinee with an unknown θ that is normally distributed.

Table 2.2 shows an example of the P_{jk} for examinees who respond with scores 0, 3, and 4 to a constructed response item which was scaled using a graded response model (Stone, 2000). From the table we see that an examinee's contribution to the item fit table is distributed over multiple θ levels, rather than restricting the contribution to a single cell based on a point estimate of θ. The sum of the probabilities for any one examinee is approximately equal to 1. If the point estimate is used to ability, then the estimated ability is -0.21, 0.63 and 1.90 for students with a score of 0, 3, and 4 respectively.

Table 2.2
Posterior Probability Distribution for Three Students Responding with Scores of 0, 3, and 4 to an Item

θ	0	1	2	3	4
QPT 1 –4.00					
QPT 2 –3.58					
QPT 3 –3.16					
QPT 4 –2.74					
QPT 5 –2.32					
QPT 6 –1.90	0.00				
QPT 7 –1.47	0.02				
QPT 8 –1.05	0.10				
QPT 9 –0.63	0.27			0.00	
QPT 10 –0.21	**0.35**			0.03	
QPT 11 0.21	0.21			0.19	
QPT 12 +0.63	0.05			**0.41**	0.00
QPT 13 +1.05	0.01			0.29	0.05
QPT 14 +1.47	0.00			0.07	0.23
QPT 15 +1.90				0.00	**0.37**
QPT 16 +2.32					0.25
QPT 17 +2.74					0.08
QPT 18 +3.16					0.01
QPT 19 +3.58					0.00
QPT 20 +4.00					

The variability of the posterior probability distribution is dependent on the precision with which ability is estimated. For longer tests, the ability is measured more precisely, and the posterior probability distribution for an examinee will be concentrated over a small range of abilities. For less precise ability estimates, the posterior distribution will be spread out over a wide range of ability.

The sum of every P_{jk} for an item across all examinees provides a pseudo-observed score distribution, which contains the number of examinees at each score level j and ability level k. Model-based predictions are calculated based on the θ_k for the subgroup and estimated item parameters. Then Pearson (χ^{*2}) and Likelihood Ratio (G^{*2}) goodness-of-fit statistics can be calculated by treating pseudocounts as observed frequencies and model-based predictions as expected frequencies. The expected frequencies for some ability levels may be 0 or very small, thus the goodness-of-fit statistic is calculated only on a subset of θ, such as the interval [-2, 2] which has less chance of sparseness than the ability outside this range. The zero frequencies for the observed and expected counts result in an undefined computation for a particular cell in the item fit table even with moderate sample size, and since the expected frequencies are always not equal to 0. To base the chi-square statistics on the same number of cells across the replications, a small constant (0.000001) was added to the cells of a zero observed frequency instead of delete the cell with zero observed frequency.

Usually a goodness-of-fit statistic is compared with a hypothesized Chi-square distribution. However, it is improper to assume that the distribution of the goodness-of-fit statistics described above follows a Chi-square distribution for two reasons. First, the assumption of Chi-square distribution is violated, as the pseudocounts are dependent on each other when an examinee's contribution to the table is no longer in one cell. Second, the goodness-of-fit statistic

is computed using estimated ability and item parameters. Therefore, Stone (2000b) investigated the null sampling distribution for the Pearson (χ^{*2}) and likelihood-ratio (G^{*2}) statistics by using Monte Carlo resampling methods. Resampling is a promising method for producing an approximate null distribution when the distribution of test statistic is unknown or in question. Stone used Q-Q plot to compare the sampling distribution for fit statistics with a theoretical chi-square distribution. In the comparison, the degrees of freedom for the theoretical chi-square distribution are equal to the mean of the sampling distribution. Since Q-Q plot was linear, Stone concluded that the fit statistics were distributed as a scaled chi-square distribution. To estimate the scaling factor (γ) and effective degrees of freedom (v) for each item, both mean and variance of empirical sampling distribution of the fit statistics were used.

To illustrate this method, the likelihood-ratio (G^{*2}) statistic, for example, is assumed to follow a scaled chi-square distribution ($G^{*2} \sim \gamma G^2$), where γ the scaling factor and v is the degrees of freedom (df) of chi-square distribution G^2. Given a chi-square distribution, the mean is equal to df and variance is equal to $2df$. Thus, it can be obtained that $E(G^{*2}) = \gamma E(G^2) = \gamma v$ and $var(G^{*2}) = \gamma^2 var(G^2) = 2v\gamma^2$. Since $E(G^{*2})$ and $var(G^{*2})$ can be estimated from the empirical sampling distribution, the scaling factor (γ) and effective degrees of freedom (v) can be determined from these equations. The rescaling method uses the estimated scaling factor and estimated degrees of freedom for hypothesis testing. The goodness-of-fit statistics are rescaled as G^{2*}/γ or χ^{2*}/γ and the hypothetical chi-square distribution is adjusted with $df = v\text{-}m$ (m is the number of item parameters).

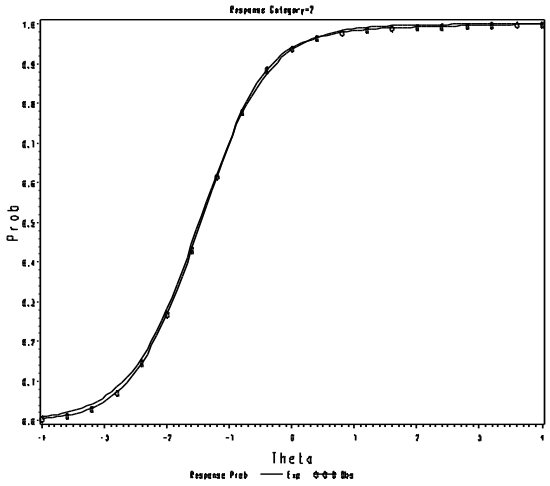
Stone and Zhang (2003) compared Orlando and Thissen's method with this rescaling method under conditions with different test length (10, 20 and 40 items) and sample size (500, 1000 and 2000 examinees). Both Orlando and Thissen's method and Stone's method found type

I error rates to be close to the nominal α regardless of test length and sample size. Empirical power was evaluated when the model used to simulate data is different from the model used to calibrate data and assess goodness-of-fit. For these two methods, the empirical power tended to increase with increased sample sizes while showing no difference with increased test lengths. It was also found that Stone's rescaling method appeared to display more power to detect model misfit than Orlando and Thissen's method, but the difference diminished as sample size increased. The performance of Orlando and Thissen's method was more seriously affected by sample size than Stone's rescaling method, and adequate power to detect modeled misfit was observed only for sample size equal to 2000 when the simulated model was 2P/3P and calibrated model was 1P. Stone's method achieved adequate power even when sample size was small. Both Orlando and Thissen's method and Stone's rescaling method lacked the power in detecting misfit when the simulated model was 3P and calibrating model was 2P. Empirical power was also evaluated by introducing misfit for a subset of items (slope parameter altered by .5 and threshold parameter altered by .25). The results indicated Stone's rescaling method displayed more power to detect model misfit than Orlando and Thissen's method.

The advantage of Stone's approach is that it accounts for uncertainty in ability estimation and allows examination of item fit when ability estimates are not precise. In addition, the rescaling method is easy to implement. However, there exits some problems associated with the adjustment of degrees of freedom in the rescaling method. The adjustment of degrees of freedom is the same regardless of the sample size and does not consider differences in precision associated with different sample sizes. For example, Figure 2.5 presents the empirical ICC based on pseudo-counts and the model-based ICC for a dichotomously scored item (a=1.80; b=-1.48) from a sample size of 1832. In the figure, the overlap of the two plots indicates consistency

between the model and the observed item response. Using the rescaling method, however, the item was determined to be misfitting. This fact may be caused by over-corrections in the degrees of freedom for the use of estimated item parameters.

Figure 2.5: Empirical and Model-based ICC for an Item (a = 1.80; b = -1.48)



2.2.4.3 Beaton Fit Indices

Residual analysis can also be used to assess model-data-fit. A residual is the difference between actual item performance for a subgroup of examinees and the subgroup's expected item performance. The traditional residual analysis approach still involves choosing an IRT model, estimating item and ability parameters, predicting the performance of various ability subgroups based on the chosen IRT model, and finally comparing the expected score distribution with the observed score distribution by using residual or standardized residual plots (Hambleton, 1990). Since the traditional residual analysis is a graphical approach, it does not depend on statistical

tests to assess goodness-of-fit. The similarity of traditional residual analysis and traditional goodness-of-fit statistics is that they both classify examinees into ability subgroups based on each examinee's estimated ability.

Beaton (2003) proposed a method for assessing goodness-of-fit by computing the fit statistics based on residuals for each examinee. Beaton's method calculates residuals for each examinee based on each examinee's Bayesian posterior ability distribution, and avoids classifying examinees into ability subgroups.

Beaton fit indices include the standardized mean residuals (the MR statistic) and the standardized mean squared residuals (the MSR statistic). To account for uncertainty in ability estimation, these two standardized residuals are calculated five times for each examinee, once for each plausible ability value. The plausible ability values are sampled from the posterior ability distribution for an examinee.

More specifically, Beaton fit indices are calculated as follows:

$$MR_i = \frac{\sum_{j=1}^{N}\sum_{a=1}^{5}\frac{(x_{ij} - E_{ija})}{\sqrt{Var_{ija}}}}{5N}, \qquad MSR_i = \frac{\sum_{j=1}^{N}\sum_{a=1}^{5}\frac{(x_{ij} - E_{ija})^2}{Var_{ija}}}{5N},$$

where MR_i is the standardized mean residuals across all examinees for item i,

MSR_i is the standardized mean squared residuals across all examinees for item i,

$j = 1, \ldots, N$ is the index of examinees,

N is the total number of examinees,

x_{ij} is the observed score of examinee j on item i,

E_{ija} is the expected score of examinee j on item i for plausible ability a, and

Var_{ij} is the variance of examinee j's score on item i.

In Beaton fit statistics, the expected score (E) is the sum of all weighted possible scores an examinee could get on an item, and the weight is the corresponding model-based probability for each possible score. The model based probabilities are calculated based on estimated item parameters and a plausible ability value sampled from the examinee's Bayesian posterior ability distribution. Thus, the expected score of examinee *j* on item *i* for plausible ability *a* is calculated as follows:

$$E = \sum_{k=0}^{K} k p_{ijk} ,$$

where $k = 0, 1, \ldots, K$ is the possible score that examinee *j* gets on item *i*, and p_{ijk} is the probability that examinee *j* gets score *k* on item *i*.

The variance of examinee *j*'s score on item *i* for plausible ability *a* is computed as follows:

$$Var = \sum_{k=0}^{K} k^2 p_{ijk} - E^2$$

For dichotomous items, the possible scores that an examinee can get are 0 and 1. Using estimated item parameters and an examinee's plausible ability, dichotomous IRT models produce model-based probabilities p_{ij0} and p_{ij1} for examinee *j* to get a score of 0 or 1 on item *i*, respectively. Therefore, the sum of p_{ij0} and p_{ij1} must be 1. The expected score of examinee *j* on item *i* for plausible ability *a* is calculated as follows:

$$E = \sum_{k=0}^{K} k p_{ijk} = 0 \times p_{ij0} + 1 \times p_{ij1} = p_{ij}$$

where p_{ij} is the probability of a correct response for a dichotomous item *i*.

The variance of examinee *j*'s score on item *i* for plausible ability *a* is calculated as follows:

$$Var = \sum_{k=0}^{K} k^2 p_{ijk} - E^2 = 0^2 \times p_{ij0} + 1^2 \times p_{ij1} - (p_{ij1})^2 = p_{ij} \times (1 - p_{ij}) = p_{ij} \times q_{ij}$$

where $q_{ij}=1-p_{ij}$ is the probability of an incorrect response for a dichotomous item *i*.

For polytomous items, when a graded response model (GRM) has three categories, the possible scores that an examinee can get are 0, 1 and 2. Using estimated item parameters and an examinee's plausible ability, the GRM produces model-based probability p_{ij0}, p_{ij1} and p_{ij2} for examinee j to get a score of 0, 1, or 2 on item i, respectively. The sum of p_{ij0}, p_{ij1} and p_{ij2} must be 1. The expected score of examinee j on item i for plausible ability a is calculated as follows:

$$E = \sum_{k=0}^{K} k p_{ijk} = 0 \times p_{ij0} + 1 \times p_{ij1} + 2 \times p_{ij2}$$

The variance of examinee j's score on item i for plausible ability a is calculated as follows:

$$Var = \sum_{k=0}^{K} k^2 p_{ijk} - E^2 = 0^2 \times p_{ij0} + 1^2 \times p_{ij1} + 2^2 \times p_{ij2} - (E)^2$$

To assess model-data-fit, Beaton proposed to generate "perturbations" under the null hypothesis. Beaton (2003) presented that "If the model is true, then the observed data may be considered a perturbation from the underlying probabilities and randomly equivalent to 'other perturbations'. If the model does not fit the observed data well, then we would expect the magnitude of the errors to be larger."

To obtain "other perturbations", Beaton proposed using bootstrap to simulate the empirical sampling distribution of MR and MSR and to conduct the corresponding hypothesis tests. Bootstrap is a widely used resampling method that was invented by Efron (1979). Bootstrap can be implemented using Monte Carlo methods, and it can be used to assess goodness-of-fit, since it can produce a rough approximation of the unknown or uncertain null hypothesis distribution of goodness-of-fit statistics. In practice, bootstrap is a computationally intensive method and used frequently in applied statistics. Now it can be realistically implemented because of the great improvement on computer performance. Hombo (as cited in Li, 2005) outlined the common bootstrap procedure in four steps:

1. Generate replicate datasets under the null hypothesis using item parameters estimated from original dataset;
2. Re-estimate item parameters for each replicate dataset;
3. Compute value of fit statistic for each replicate dataset;
4. Compare value of fit statistic calculated from original dataset to those values from replicate datasets.

It should be noted that some simulation studies include step 2 while others do not.

In Beaton's method, if MR and MSR of the observed dataset are randomly equivalent to the MRs and MSRs of simulated datasets, it can be concluded that the observed data fits the IRT model. However, if MR and MSR of the observed data seriously deviate from the simulated data, the observed data is determined not to fit the IRT model.

In Li's (2005) dissertation, she generated 200 resamples to test model-data-fit. That is, "The study takes 200 resamples, so for each person on each item for each plausible value, there will be 200 simulated student scores." "For the 200 resamples, 200 pairs of MR and MSR can be calculated, the observed MR and MSR statistics will be calculated from the original student and then be compared with the 200 corresponding overall model-data fit statistics calculated from the resamples. The 200 resamples are independently simulated under the null hypothesis. In testing overall model-data-fit, if the null hypothesis holds, the observed MR and MSR statistics and the 200 pairs of MR and MSR from resamples are equally likely values. If the occurrence of the observed fit statistics is more than what would be expected from random fluctuation, we conclude the model does not fit the observed data. The MR statistic is signed, if the proportion of simulated MRs that are greater than the observed MR is less than .025 or greater than .975, the null hypothesis of good model-data fit is rejected. The MSR statistic is unsigned. We reject the

null hypothesis if the proportion of simulated MSRs that are greater than the observed MSR is less than .05."

Both Beaton's method and Stone's method compute fit statistics by using estimated abilities based on posterior distribution. The posterior distribution combines the information from the prior distribution of ability and the likelihood function. Both methods take into account the imprecision of ability estimation. In Beaton's method, the different plausible ability values reflect the uncertainty in the abilities. In Stone's method, the posterior expectations at each discrete ability level are used to reflect the imprecision of ability estimation. As all goodness-of-fit methods can be expressed in a two-way contingency table, Beaton's method uses rows to represent examinees, while Stone's method uses rows to represent ability subgroups. In Stone's method, subgroups are formed by dividing ability continuum into discrete intervals.

The advantages of Beaton's method are that it does not depend on a specific asymptotic null distribution to assess goodness-of-fit and does not depend on a specific methodology to group examinees. However, since Beaton uses resampling methods to simulate the null distribution, the computation of these statistics is time consuming.

CHAPTER 3

METHODOLOGY

Monte Carlo resampling methods were used to evaluate the performance of Beaton fit indices. Monte Carlo resampling can provide an alternative solution to statistical problems when an analytical solution is not available. The theoretical distribution of Beaton fit indices is unknown, so Monte Carlo resampling methods can be useful to generate the empirical null distribution and test the hypothesis about model-data-fit.

The methodology is presented in this chapter. It consists of the manipulated factors under study, the procedures for data simulation, the procedures for Monte Carlo resampling to test model-data-fit, the methodology for evaluating the Beaton fit indices and an analysis plan. There were several objectives for this simulation study. Firstly, the efficacy of the Beaton fit indices were evaluated by investigating type I error rates, empirical power and the distribution of Beaton fit indices. The testing conditions (e.g., sample size, test length) under which the procedure is appropriate was also of interest. Given that a Monte Carlo resampling procedure was proposed for hypothesis testing, it was also important to evaluate how many Monte Carlo resamples under the null hypothesis are enough to adequately determine the critical values in the tails of the sampling distribution.

3.1 Factors under Study.

Three factors were crossed in this study: test length (12, 24 and 36 items), sample size (500, 1000 and 2000) and Monte Carlo resample size (100 and 200). These three manipulated factors are assumed to have an important influence on goodness-of-fit statistics.

Test length is related to the precision of ability estimation. Different test lengths will introduce variability in the accuracy of ability estimates for each examinee. The selection of different test lengths allows the procedures in this study to be investigated across a wide range of conditions. A test length of 12 items was used to conform to a typical performance assessment and an assessment incorporating matrix sampling methods (e.g. NAEP). The test lengths of 24 and 36 items were selected to conform to conditions under which ability can be measured more precisely. Sample size is related to the precision of item parameters estimation. Larger sample sizes produce item parameter estimates with greater precision. The sample sizes were selected to be larger than 500 which is the recommended sample size for accurate estimation of item parameters. The selected sample sizes of this study were chosen to provide consistent item parameter estimates and reflect sample sizes that are consistent with larger scale assessment programs (e.g., NAEP). The last manipulated factor was Monte Carlo resample size. By comparing different Monte Carlo resample sizes, the number of resamples needed for hypothesis testing can be evaluated.

These levels of test lengths and examinee sample sizes were selected for the comparison with the previous methods. These include Orlando and Thissen's method (Orlando & Thissen, 2000) and Stone's method of the fit statistics based on posterior expectations (Stone, 2003).

3.2 Item Parameters

Item parameters for this simulation study were based on an analysis of item responses from the 1994 NAEP reading administration assessment. The assessment was composed of a mix of dichotomously and polytomously scored items. A subset of 6 items was selected from the 1994 NAEP assessment and comprised of four 2-category items, one 3-category item and one 4-category item. The chosen items were similar to other items in NAEP blocks, and therefore are representative of the items in the 1994 NAEP assessment. This set of 6 items was replicated to obtain sets of 12, 24 and 36 items. This ensured consistency across the experimental conditions. Table 3.1 presents the item parameters (*a* – slope parameter, *b* – threshold parameter):

Table 3.1:
Item Parameters for the Simulation Study (6 Items)

Item	a	b_1	b_2	b_3
1	.49	-.89		
2	1.55	1.23		
3	1.10	-.36		
4	.91	-1.42		
5	2.08	.33	1.44	
6	.89	-1.26	0.93	2.89

3.3 Generating Item Responses

The item responses were simulated based on the item parameters and ability parameters from a specific distribution. Simulating the real item responses involves the following steps:

1. Randomly generate an ability parameter from a standard normal ability distribution *N(0, 1)*,

2. Calculate the probability of an examinee's response using the item parameters in Table 3.1 and ability parameter from step 1,

3. Generate a random number *(r)* from an uniform distribution *U(0, 1)*, and

4. Compare the probability in step 2 with a random number in step 3. For a dichotomous item, if the probability of a correct response for an examinee is larger than the randomly generated number, then a simulated score of *1* is assigned to the examinee on this item; otherwise, a simulated score of *0* is assigned. For the graded response model, the random number is compared with boundary category response function of the GRM. For example, suppose the random number is *r* and an item with *k* categories boundary response function *(p_1^*,..... p_k^*).* Then the simulated observed item response *x* is given as following:

$$if \;\; 1\text{-}p^*_{n-1} < r < 1\text{-}p^*_{n} \Rightarrow x = n\text{-}1 \qquad n=1, \ldots, k$$

$$else \;\; x=k$$

This procedure was used to simulate real item responses for combinations of 3 test lengths, 3 sample sizes and 2 Monte Carlo resampling sizes. For each combination, a number of replications (NR) of item responses were generated. In this study, three levels of replications (NR=100 and 200) were specified. To accurately evaluate the sampling distribution of Beaton fit indices, a fairly large number of replications (e.g., 500) may be needed. However, in terms of evaluating the use of the method in practical applications, it is important to see if a smaller number of replications are adequate to achieve the desired power and type I error rates.

3.4 Calibrating the Data

The item responses were calibrated using the computer program MULTILOG, since MULTILOG can be used to estimate the item parameters for both dichotomous and polytomous items.

3.5 Procedures for Testing the Goodness-of-fit of Beaton's Method

The steps of Monte Carlo Simulation for testing the hypothesis using Beaton fit statistics are as follows:

1) From simulated real item responses, estimate the IRT model using MULTILOG and posterior ability distribution for each examinee. Then, randomly sample 5 plausible ability values from each examinee's posterior ability distribution and form 5 sets of plausible ability values for all examinees. Finally, the MR and MSR across all examinees for simulated real item responses are obtained.
2) Given the estimates of item parameters and an assumed normal ability distribution, generate a random set of response vectors of size N (size of observed sample) under the null hypothesis; and compute the MR and MSR.
3) Repeat step 2 for R times (i.e., R is the Monte Carlo resample size) to produce an empirical sampling distribution for MR and MSR.
4) For MR, using the interval 5^{th}-95^{th}, 2.5^{th}–97.5^{th} and 0.5^{th}–99.5^{th} from the empirical sampling distribution, determine if the MR for simulated real item responses (from step 1) sits within the intervals to assess the significance of the test statistic. For MSR, using the 90^{th}, 95^{th} and 99^{th} percentiles from the empirical sampling distribution, determine if

the MSR for simulated real item responses (from step 1) is lower than the P_{90}, P_{95} or P_{99} to assess the statistical significance of the test statistic.

3.6 Evaluation of Beaton Fit Indices

This simulation study evaluated Type I error rates, empirical power for Beaton fit indices under various conditions. For each combination of test conditions, a number of replications (NR=100 and 200) which constitute the empirical distribution of Beaton fit indices were generated. Then the percentage of items identified as misfitting across the number of replications was computed. This percentage was then used to evaluate Type I error rates or empirical power rates based on the relationship between the model used to simulate the data and the model used to calibrate data.

3.6.1 Type I error rates

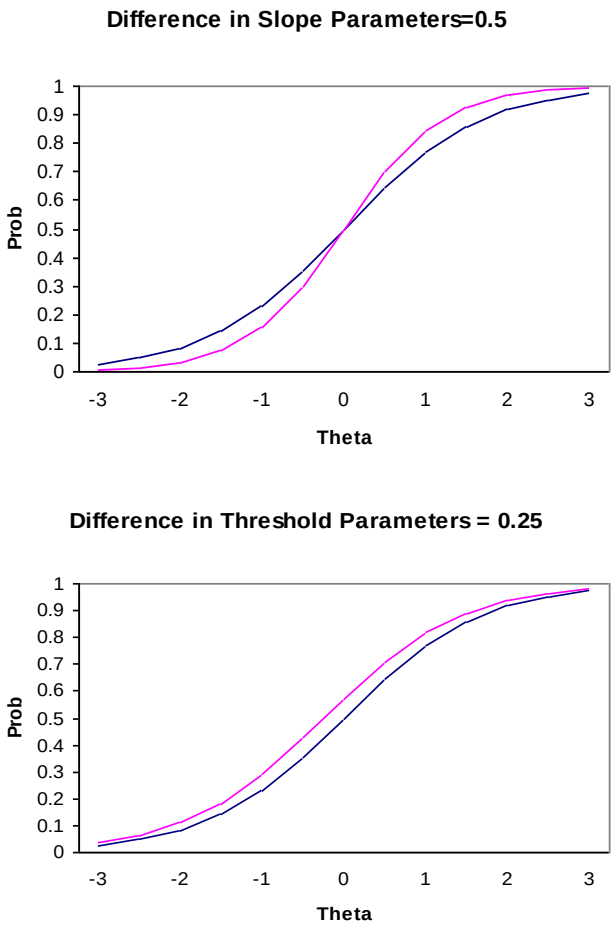
Type I error rates were evaluated when the item responses are simulated under the null hypothesis. That is, the model used to generate the item responses is the same as the model used to estimate the model parameters. The Type I error rates are the proportions of times items are found to be misfitting when the items truly fit the model. Three nominal rejection rates (α) were examined: 0.10, 0.05 and 0.01. Since this simulation study generated the empirical distribution of Beaton fit indices under the null hypothesis, the sampling distribution was also investigated.

3.6.2 Empirical power

Empirical power was evaluated as the percentage of misfit when the null hypothesis is false. The model misfit conditions were introduced in two ways. For one case, the null

hypothesis was false for all the test items. Thus, the model (initial model) used to simulate item responses is different from that used to scale and assess goodness-of-fit. For this case, the slope parameters in the initial model were not required to be equal, but they were constrained to be equal in assessing goodness-of-fit. Since this simulation included both the 2P and GRM models, for the initial 2P model, the scale model was 1P, and for the initial GRM, the scale model had the same slope parameter (*a*) as the scale model for the 2P model. This constraint was implemented using MULTILOG. For a second case, the null hypothesis was false for a subset of items. For small subset of item parameters, a false null hypothesis was imposed for computing goodness-of-fit statistics. This type of misfit was introduced by incorporating a small to moderate difference in some of the item parameters used to simulate item responses and on the estimated item parameters used to calculate fit statistics. For example, while item responses were simulated with a slope parameter of 1.2 for an item, the fit statistic is computed under an alternative hypothesis with a slope parameter of 0.7. In this study, a 0.5 difference for a slope parameter and a 0.25 difference for a threshold parameter were manipulated. For instance, the slope parameter for the first item (**a=-0.49**, b=-0.89) and the first threshold parameter for the last item (a=0.89, $\mathbf{b_1}$**=-1.26**, b_2=0.93, b_3=2.89) were altered. Figure 3.1 shows an example of item characteristic curves (ICCs) that reflect the effects of altering the item parameters.

Figure 3.1: ICCs Illustrate the Effects of Altering Item Parameters



3.7 Analysis Plan

The sampling distribution of Beaton fit indices across the number of replications were evaluated from two perspectives: First, the mean and variance for the items across 1000 replications was examined. Second, Quantile-Quantile (Q-Q) plots were used to explore whether the sampling distribution followed a theoretical distribution. When the empirical and theoretical distributions match, Q-Q plots will be linear close to line $y=x$ (x-horizontal axis, y-vertical axis). When the two distributions do not match, Q-Q plots provide information showing how an empirical distribution might deviate from a theoretical distribution. For example, when Q-Q plots

are linear but do not fall on the line $y=x$, this suggests that the empirical distribution is a member of the theoretical distribution family. When Q-Q plots are not linear at all, the empirical distribution is different from the theoretical distribution.

Although the investigation of the empirical distribution of Beaton fit indices is of value, the more important investigation is to evaluate their performance under the proposed hypothesis testing procedure. The reason is that Beaton's method does not rely on a theoretical distribution for hypothesis testing or to assess model-data-fit. Rather, a Monte Carlo resampling-based method is proposed. Therefore, to evaluate the efficacy of the procedure it is important to evalute Type I error rates and empirical power under a variety of testing conditions.

In the study, the Type I error rate is the percentage of false rejections across the number of replications (NR=100 and 200). The empirical power rate is the percentage of correct rejections across the number of replications. To investigate the performance and effect of independent factors (test length, sample size and Monte Carlo resample size), tabular and graphical summaries of Type I error rates and empirical power were used. To examine whether a factor was significant or not, an analysis of variance (ANOVA) test was conducted. For the ANOVA test, the independent variables were test length, sample size, Monte Carlo resample size and number of replications, and the dependent variable was empirical power rates at three α levels.

CHAPTER 4

RESULTS

The purpose of this study was to evaluate the statistical properties of Beaton's MR and MSR statistics. In this chapter, the statistical results of Beaton's MR and MSR statistics are presented for combinations of different simulated factors. The different simulated factors were test length, sample size, Monte Carlo resample size and number of replications. The statistical results are presented separately for three parts.

First, the sampling distributions of Beaton's MR and MSR statistics were investigated. To investigate the sampling distribution, Quantile-Quantile (Q-Q) plots of Beaton's MR and MSR statistics versus an assumed theoretical normal distribution were constructed. Tables summarizing the means and standard deviations of Beaton's MR and MSR statistics were also evaluated. Second, Type I error rates for Beaton's MR and MSR statistics were studied to examine the behavior of the statistic in the tails of the sampling distribution. Finally, the statistical power of Beaton's MR and MSR statistics in detecting misfit was investigated under different combinations of simulated factors. This part investigated whether adequate power existed to detect misfit and whether the empirical power was affected by different factors. To evaluate the effects of the different factors, Analysis of Variance (ANOVA) tests were performed.

4.1 Sampling distribution for Beaton's MR and MSR

The empirical sampling distributions of Beaton's MR and MSR statistics for each item were evaluated in this study. To obtain the sampling distributions, one thousand replications were generated for a 12 item test with different sample sizes (500, 1000 and 2000). The general procedures for generating the sampling distributions were as follows: (a) Item response data were simulated using the original item parameters and a randomly generated ability θ; (b) The simulated item response data were calibrated using MULTILOG; (c) Each examinee's posterior ability distribution based on the estimated item parameters from step b was estimated, from which a random sample of five plausible abilities was obtained; (d) Beaton's MR and MSR statistics were then calculated using the item parameter estimates from step b and the plausible abilities from step c; and (e) Steps a through d were repeated 1000 times.

To evaluate the sampling distribution, the empirical distribution for each test item was compared with a theoretical distribution. The comparison was conducted through a quantile-quantile (Q-Q) plot. A Q-Q plot involves plotting the empirical quantiles from the observed data against the corresponding quantiles from a theoretical distribution. The expected pattern of the Q-Q plot, should the data fit the distribution, is a straight line with intercepts of 0 and slope of 1. Any distributional differences will appear as deviations from this straight line pattern. Q-Q plots that are linear with slopes different from 1 indicate that the empirical and theoretical distributions are from the same family of distributions but differ in dispersion. Q-Q plots that are linear with intercepts different from 0 indicate that the empirical and theoretical distributions differ in location.

4.1.1 Sampling distribution for Beaton's MR

There is no previous research investigating the sampling distribution of mean standardized residual or mean squared standardized residual. In this study, the theoretical distribution was assumed based on the central limit theorem. The Central Limit Theorem states that, when sample size is large enough, the sampling distribution of the mean of the observation is well approximated by a normal distribution, even when the population distribution is not itself normal. Therefore, the theoretical distribution of Beaton's MR and MSR statistics was assumed to be a normal distribution. Since Beaton's MR statistic is the mean of standardized residuals across the five plausible ability values ($MR_i = \frac{\sum_{j=1}^{N}\sum_{a=1}^{5}\frac{(x_{ij} - E_{ija})}{\sqrt{Var_{ija}}}}{5N}$) and the distribution of standardized residual is the standard normal distribution N (0, 1), the sampling distribution of MR may be hypothesized to follow a normal distribution with mean = 0 and variance = 1/(5N) (N is the sample size). In this study, three sample sizes were manipulated, so hypothesized theoretical distributions for Beaton's MR fit statistic are N(0, 0.0004), N(0, 0.0002) and N(0, 0.0001) with sample size equal to 500, 1000 and 2000 respectively.

Figures 4.1-4.6 present the Q-Q plots for the MR statistic (sample size N=500) versus a theoretical normal distribution N (0, 0.0004). Since six items were replicated to generate the 12 item tests in this study, Q-Q plots of the first six items are presented in the figures. A diagonal straight line is superimposed on the Q-Q plots to indicate the expected plot given similarity between the empirical and theoretical distributions for the fit statistics.

Figure 4.1
Normal Q-Q plot of MR Statistic for Item 1
(α = .49, *b* = -.89)

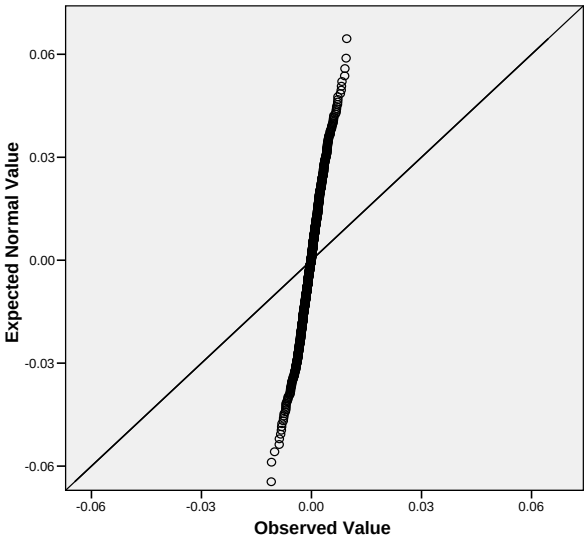


Figure 4.2
Normal Q-Q plot of MR Statistic for Item 2
(α = 1.55, *b* = 1.23)

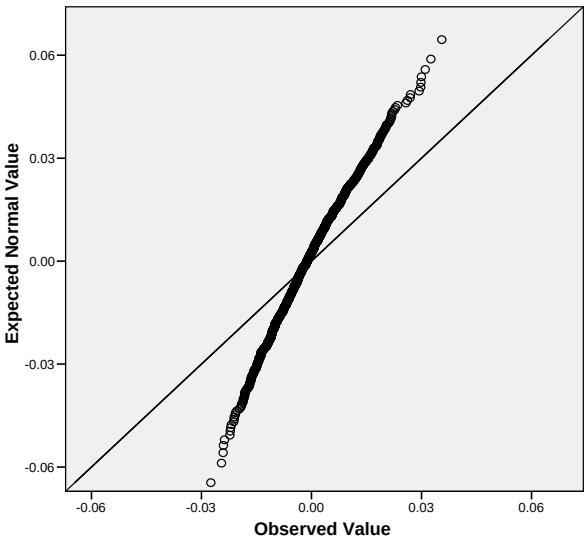


Figure 4.3
Normal Q-Q plot of MR Statistic for Item 3
(α = 1.10, *b* = -.36)

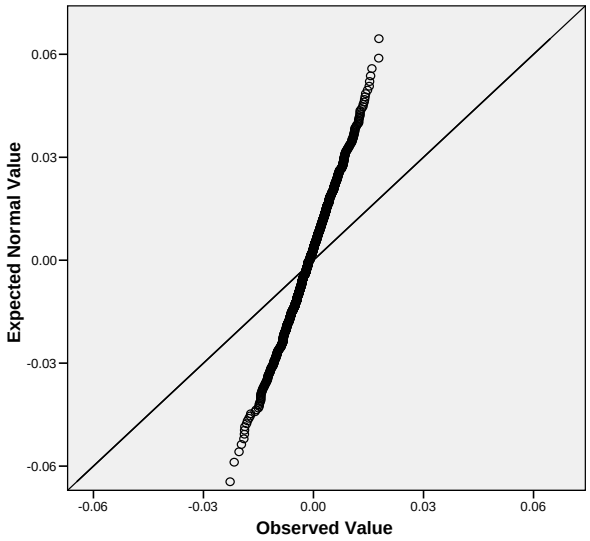


Figure 4.4
Normal Q-Q plot of MR Statistic for Item 4
(α = .91, *b* = -1.42)

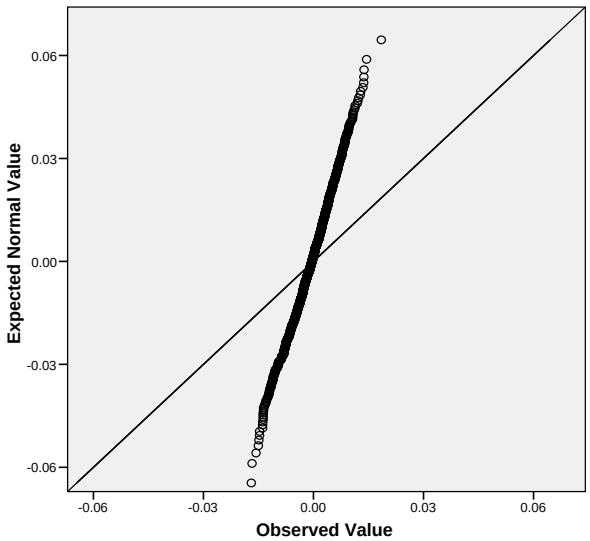


Figure 4.5
Normal Q-Q plot of MR Statistic for Item 5
(α = 2.08, b_1 = .33, b_2 = *1.44)*

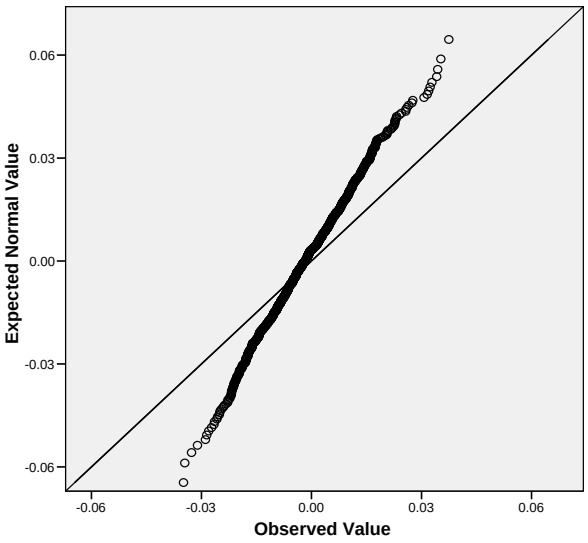
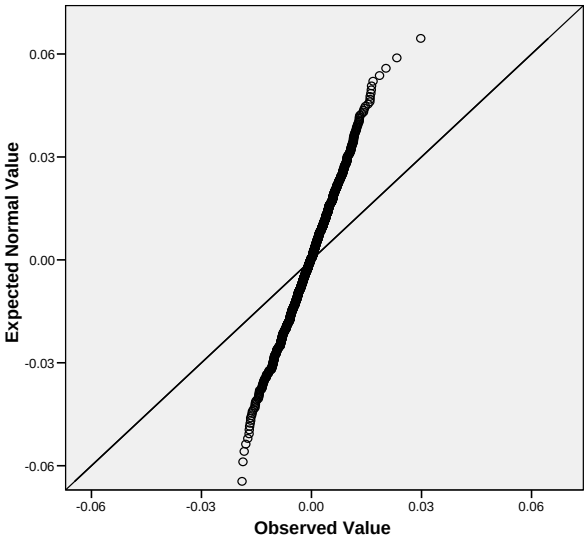


Figure 4.6
Normal Q-Q plot of MR Statistic for Item 6
(α = .89, b_1 = -1.26, b_2 = *.93*, b_3 = *2.89)*



As can be seen, Q-Q plots for Beaton's MR statistic showed similar patterns across different test items. The Q-Q plots were all linear with little deviation at the lower and upper tail. However, the deviation in the plot of fit statistics from the diagonal straight line indicated that the empirical sampling distributions belonged to the family of normal distributions but differed from a N(0,0.0004).

Table 4.1 presents the means, standard deviations, and skewness and kurtosis statistics of the first six items for Beaton's MR statistic for the 12 items test condition with different sample sizes (500, 1000 and 2000). From table 4.1, the skewness and kurtosis statistics indicated the sampling distributions were approximately normal, the estimated means and standard deviations were all close to zero indicating very little variability across replications. Given no basis for a theoretical sampling distribution to test the hypothesis of model-data-fit, a Monte Carlo resampling method would be required to test the hypothesis of model-data-fit for Beaton's MR statistic.

Table 4.1
Means, Standard Deviations, Skewness and Kurtosis Statistics for MR (12 Items Test)

Sample size	Item #	Mean	Standard deviation	Skewness	Kurtosis
500	**1**	-0.00034	0.00285	-0.10614	0.57140
	2	-0.00084	0.00924	0.25741	0.07056
	3	-0.00095	0.00669	-0.00923	0.39891
	4	-0.00060	0.00538	-0.29747	0.32440
	5	-0.00214	0.01145	-0.08057	-0.05778
	6	-0.00037	0.00694	-0.07823	0.27261
1000	**1**	-0.00019	0.00189	-0.00094	0.53343
	2	-0.00067	0.00691	0.13387	0.31217
	3	-0.00071	0.00478	0.04055	-0.12788
	4	-0.00051	0.00395	0.08159	0.04206
	5	-0.00174	0.00786	0.11969	-0.17992
	6	-0.00033	0.00483	-0.03540	-0.34375
2000	**1**	-0.00016	0.00135	-0.06539	0.20566
	2	-0.00053	0.00471	0.05855	-0.05985
	3	-0.00046	0.00331	0.01731	0.16584
	4	-0.00031	0.00276	-0.03529	0.00313
	5	-0.00146	0.00562	-0.00505	0.25651
	6	-0.00026	0.00348	-0.07534	0.11491

4.1.2 Sampling distribution for Beaton's MSR

The sampling distributions of Beaton's MSR statistic were also examined. Beaton's MSR statistic is the mean of squared standardized residual across the five plausible ability values. From the Central Limit Theorem, the sampling distribution may be a normal distribution. To specify the theoretical normal distribution, the means and variances for the test items were investigated. Table 4.2 presents the means, standard deviations, and skewness and kurtosis statistics for Beaton's MSR statistic for the first six items under the conditions of 12 item tests with different sample sizes (500, 1000 and 2000). From table 4.2, the means for the test items were all approximately 1. Thus, the theoretical distribution of MSR statistic was assumed to be normal distribution (N (1, 0.0004)) which has the same variance as the theoretical distribution for MR statistic. This is reasonable since the MSR statistic reflects the square of the MR statistic, and the MR statistic was hypothesized to be a z-statistic. Since z^2 would be assumed to follow a chi-square distribution with 1 df, the mean would be hypothesized to be 1.

Figures 4.7-4.12 present Q-Q plots of empirical distributions of Beaton's MSR statistic for 12 item tests with a sample size of N=500 versus the theoretical normal distribution (N(1, 0.0004)) for the first six test items.

Table 4.2
Means, Standard Deviations, Skewness and Kurtosis Statistics for MSR (12 Items Test)

Sample size	Item #	Mean	Standard deviation	Skewness	Kurtosis
500	**1**	1.00024	0.00667	0.11503	1.23237
	2	0.99083	0.09073	1.40871	5.24469
	3	0.99896	0.02237	0.33242	1.36513
	4	1.00028	0.02867	0.54388	2.04654
	5	0.98326	0.09315	3.37739	29.49409
	6	0.99934	0.01589	0.10281	0.30945
1000	**1**	1.00028	0.00481	0.31152	1.21901
	2	0.99114	0.06588	0.98443	2.74351
	3	0.99973	0.01468	0.06535	0.59271
	4	1.00011	0.02076	0.32612	0.85975
	5	0.98230	0.05927	1.67274	7.01803
	6	0.99930	0.01106	0.03223	0.03883
2000	**1**	1.00003	0.00300	0.01532	0.29742
	2	0.99322	0.04301	0.56944	1.97086
	3	0.99977	0.01063	0.24559	0.48526
	4	1.00088	0.01434	0.20446	0.17695
	5	0.98526	0.04516	1.63538	8.35241
	6	0.99917	0.00750	0.08973	0.04797

Figure 4.7
Normal Q-Q plot of MSR Statistic for Item 1
(α = .49, *b* = -.89)

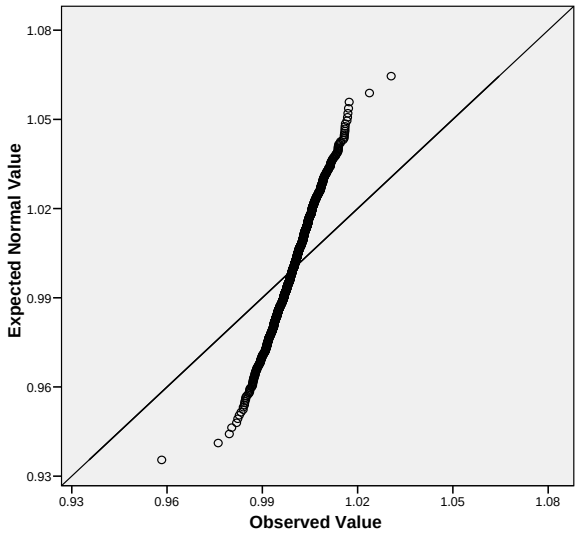


Figure 4.8
Normal Q-Q plot of MSR Statistic for Item 2
(α = 1.55, *b* = 1.23)

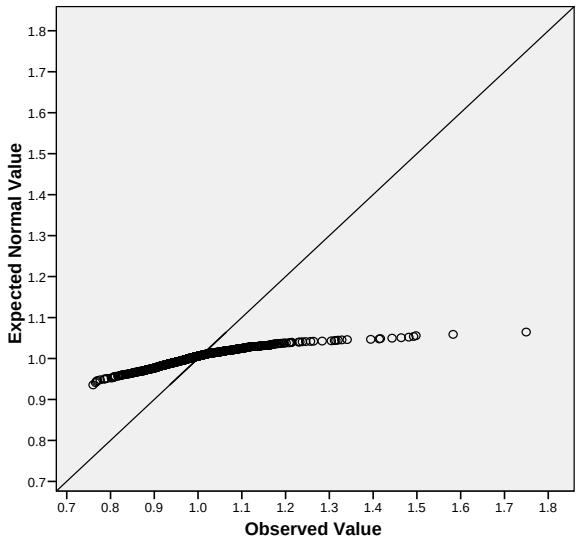


Figure 4.9
Normal Q-Q plot of MSR Statistic for Item 3
(α = 1.10, *b* = -.36)

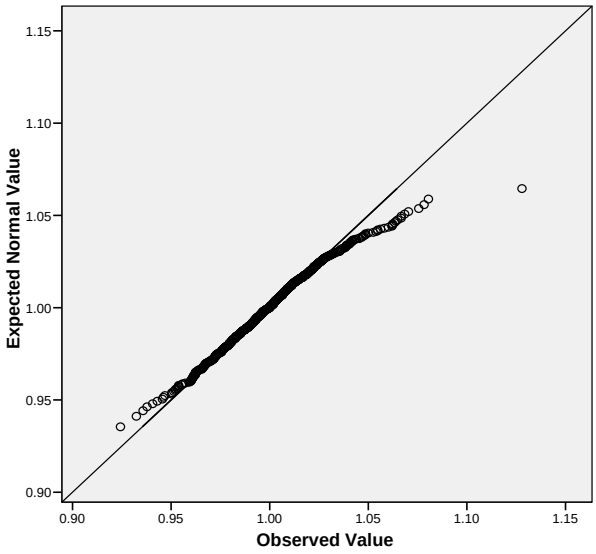


Figure 4.10
Normal Q-Q plot of MSR Statistic for Item 4
(α = .91, *b* = -1.42)

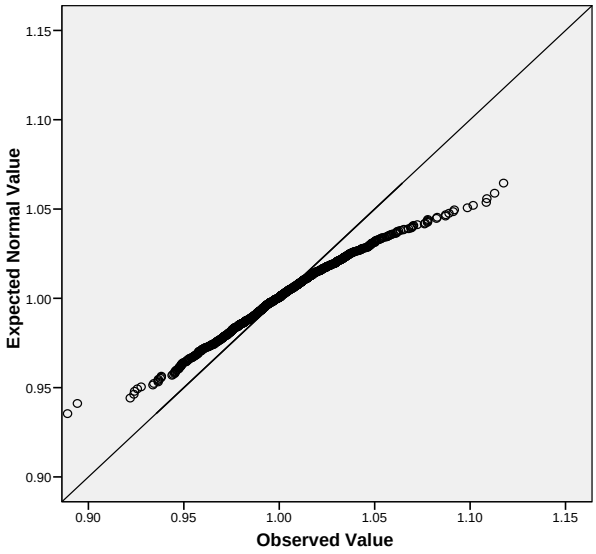


Figure 4.11
Normal Q-Q plot of MSR Statistic for Item 5
(α = 2.08, b_1 = .33, b_2 = *1.44)*

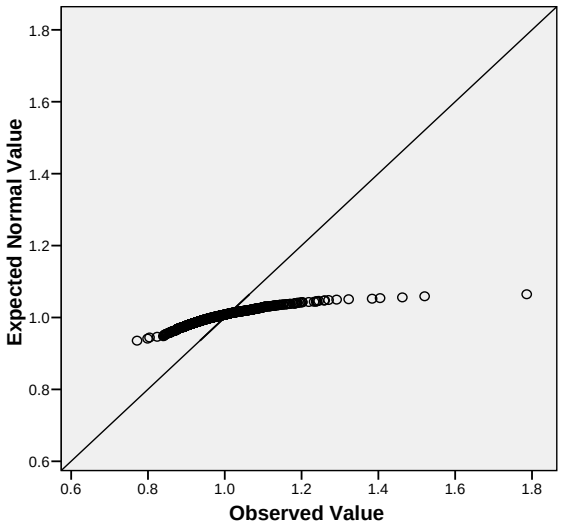
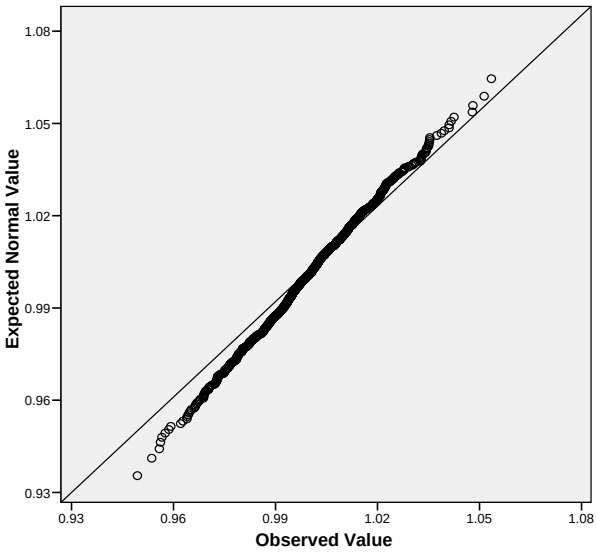


Figure 4.12
Normal Q-Q plot of MSR Statistic for Item 6
(α = .89, b_1 = -1.26, b_2 = *.93*, b_3 = *2.89)*



As can be seen, Q-Q plots were all linear but deviated from the expect pattern. Therefore, the sampling distributions of MSR statistic are belonging to the family of normal distribution. From the table of means, standard deviations, and skewness and kurtosis statistics for Beaton's, it can be seen that most of the skewness and kurtosis statistics for Beaton's MSR statistic were larger than those for Beaton's MR statistic, indicating more variability from the theoretical normal distribution for Beaton's MSR statistic than the MR statistic. The means for different items were all around 1 and the variances for different items were small and varied more than the MR variances. Therefore, as was the case with Beaton's MR statistic, a Monte Carlo resampling method would be required to test the hypothesis of model-data-fit for Beaton's MSR statistic since there is no basis for a theoretical sampling distribution to test the hypothesis of model-data-fit.

4.2 Type I Error rates

Type I error rates for Beaton's MR and MSR statistics were examined when the null hypothesis for all the test items was true (i.e., items fit the model used to simulate item responses). In this study, Monte Carlo resampling methods were used to generate the sampling distribution and test the hypothesis of model data fit at three nominal α levels (0.01, 0.05 and 0.10). The general procedures for investigation of Type I error rates involved: 1) item responses were simulated based on original item parameter, and then the item responses were calibrated using MULTILOG and each examinee's posterior ability distribution was obtained. After that, the observed MR and MSR were obtained based on the item responses, estimated item parameters and five plausible ability values from each examinee's posterior ability distribution.

2) Beaton's MR and MSR sampling distributions were obtained by repeating step 1 except that the item responses were simulated based on the estimated item parameters. 3) Model-fit decisions (e.g., Reject or accept the null hypothesis of model-data-fit) were made by comparing the observed MR and MSR statistics from step 1 with cut values derived from the MR and MSR sampling distributions obtained in step 2 at α equal to .10, .05 and .01. Since the MR statistic is signed, α/2 and 1- α/2 percentiles from the MR sampling distribution were used as cut values. If the observed MR statistic is less than α/2 or larger than 1- α/2 percentile, then the null hypothesis of model-data-fit was rejected. Since the MSR statistic is unsigned, (1- α) percentiles from the MSR sampling distribution were used as cut values. If the observed MR statistic was larger than the (1- α) percentile, then the null hypothesis of model-data-fit was rejected. 4) Steps 1 to 3 were repeated 100 or 200 times, and the average percentage of false rejections across the number of replications was calculated across all the test items. The expected false rejections for the test items were equal to the nominal rate α. To account for sampling error around the expectations, 95% confidence intervals for the expected proportions of Type I error rates were considered. For example, across the number of replications, Type I error rates of 0-3, 1-9, and 4-16 were expected for NR=100 and α=0.01, α=0.05, α=0.10, respectively.

Tables 4.3-4.5 present the Type I error rates for Beaton's MR and MSR statistics using the above resampling procedure under combinations of manipulated factors (test length, number of replications, sample size and Monte Carlo resample size). Entries in the tables represent the percentage of times across the number of replications that item misfit was detected when H_0 was true for all the test items. Note that the tables reflect a summary across the set of test items, that is, percentages were averaged over all test items.

Table 4.3
Type I Error Rates for Beaton's Fit Statistics (12 items)

Number of replications (r)	Monte Carlo sample size (R)	Sample Size (N)	MR			MSR		
			α=1%	α=5%	α=10%	α=1%	α=5%	α=10%
100	**100**	**500**	2.67	7.67	13.5	0.75	4.67	9.83
		1000	1.25	4.33	9.41	1.5	4.25	8.83
		2000	1.27	5.41	10.41	1	5.58	11.16
	200	**500**	1.67	5.91	10.58	0.58	4.58	9.75
		1000	1	6.16	11.25	0.5	5	10.16
		2000	1.41	4.58	9.41	1	5.41	11.58
200	**100**	**500**	1.54	4.87	10	0.75	4.75	9.45
		1000	1.21	4.37	9.12	1.37	4.79	9.58
		2000	1.91	6.37	11.87	0.66	5.62	10.83
	200	**500**	1.16	4.62	9.21	0.62	4.62	9.71
		1000	1.08	5.5	10.29	0.62	4.83	10.16
		2000	1.41	5.16	10.75	0.87	5	10.33

- *Percentages reflect averages over the set of items across the number of replications.*

Table 4.4
Type I Error Rates for Beaton's Fit Statistics (24 items)

Number of replications (r)	Monte Carlo samples (R)	Sample Size (N)	MR			MSR		
			α=1%	**α=5%**	**α=10%**	**α=1%**	**α=5%**	**α=10%**
100	**100**	**500**	1.91	5.41	10.08	1.04	5.12	11.16
		1000	1.37	5.87	11.04	1.04	5.25	10.29
		2000	2	6.08	11.25	1.08	5.29	10.12
	200	**500**	0.95	4.79	10.16	0.95	4.75	10.25
		1000	1.29	4.83	8.71	0.71	4.83	9.95
		2000	0.66	4.87	9.71	1.04	4.79	10
200	**100**	**500**	1.66	4.77	9.41	0.83	4.87	10.79
		1000	1.37	5.52	10.43	1	5.58	10.83
		2000	1.89	5.31	9.93	1	4.89	9.45
	200	**500**	1.41	5.35	11.02	0.81	4.83	10.29
		1000	1.18	4.60	8.83	0.77	4.93	9.77
		2000	1.33	5.5	10.45	1.04	5	9.60

- *Percentages reflect averages over the set of items across the number of replications.*

Table 4.5
Type I Error Rates for Beaton's Fit Statistics (36 items)

Number of replications (r)	Monte Carlo samples (R)	Sample Size (N)	MR			MSR		
			α=1%	α=5%	α=10%	α=1%	α=5%	α=10%
100	**100**	**500**	2.02	6.08	10.88	0.77	5.02	10.81
		1000	1.83	6.19	10.72	1.42	5.33	10.47
		2000	1.67	5.55	10.97	0.78	5.11	9.78
	200	**500**	0.97	5	9.13	0.55	4.02	9.44
		1000	0.69	4	8.03	0.75	4.86	10
		2000	0.77	3.69	8.52	1.02	4.05	8.83
200	**100**	**500**	1.72	5.33	10.16	0.77	4.72	10.44
		1000	1.79	6.06	10.91	1.25	5.25	10.34
		2000	1.97	5.58	10.61	0.88	5.16	9.95
	200	**500**	0.93	4.62	9.12	0.73	4.26	9.61
		1000	0.97	4.98	9.66	0.72	4.77	10
		2000	0.94	4.26	9.19	1.16	4.83	9.79

- *Percentages reflect averages over the set of items across the number of replications.*

As can be seen from the tables, Beaton's MR and MSR statistics provided similar results regarding Type I error rates. As would be expected given the resampling methods that were used, nominal Type I error rates were observed for both Beaton's MR and MSR statistics across different combinations of simulated factors. For example, for α=.05, the Type I error rates for Beaton's MR statistic was 5.87 and type I error rates for Beaton's MSR statistic was 5.25 under the combination of different factors (test length: 24; sample size=1000; Monte Carlo sample size=100 and number of replications=100).With the consideration of the sampling error, the entries of type I error rates in tables 4.3 - 4.5 were all within the 95% confidence intervals of the expected number of false rejections. There was not much difference in Type I error rates for Beaton's MR and MSR statistics for different sample sizes as well for different test lengths, Monte Carlo resample sizes and number of replications. A small Monte Carlo resample size (e.g. R=100) appeared adequate to support nominal type I error rates. This is a useful result since fewer Monte Carlo samples implies less computer time to perform hypothesis testing.

4.3 Empirical power

Empirical power was investigated when some or all items that were manipulated did not fit the underlying model. In this simulation study, model misfit was introduced in two different ways: (1) The model used to simulate data was different from the model used to evaluate goodness-of-fit, that is, when H_0 was false for the entire test; (2) The parameter estimates for a subset of items used to calculate the fit statistics were altered from the parameters used to generate the item responses, that is, the null hypothesis was false for a subset of test items. To investigate empirical power, the percentages of correct rejections across the number of replications for the combinations of different factors were investigated at three α levels (0.01, 0.05 and 0.10).

4.3.1 Empirical power under the condition that H_0 was false for all test items

The first type of model misfit was introduced by simulating data with different slope parameters but scaling and evaluating fit with a constant slope parameter. More specifically, in this study, item responses were simulated using a 2P model and GRM. If the 2P model was used to simulate item responses, a 1P model was the model used to estimate item parameters and evaluate model-data-fit. If the GRM was used to simulate item responses, a GRM model with the same slope parameter as 1P model (the calibrating model for 2P) was used to estimate item parameters and evaluate model-data-fit.

Tables 4.6 – 4.8 summarize the empirical power rates for Beaton's MR and MSR statistics using resampling procedure under conditions that H_0 was false for all the test items. The tables provide average percentages of correct rejections (empirical power rates) across all the test items for the combinations of three test lengths (12, 24 and 36), three sample sizes (500,1000 and 2000) ,two Monte Carlo resample sizes (100 and 200) and two number of replications (100 and 200).

Table 4.6
Empirical Power Rates for Beaton's Fit Statistics (12 items)

Number of Replications (r)	Monte Carlo samples(R)	Sample Size (N)	MR			MSR		
			α=1%	α=5%	α=10%	α=1%	α=5%	α=10%
100	100	500	26.50	37.83	45.33	17.91	24.16	29.41
		1000	42.16	53.58	58.50	24.58	30.58	34.41
		2000	55.00	63.91	68.66	27.91	36.16	41.00
	200	500	24.75	39.16	47.83	16.67	24.91	30.00
		1000	40.00	52.75	57.41	23.50	30.25	34.16
		2000	55.66	66.08	71.41	27.58	35.33	40.08
200	100	500	26.08	37.45	44.79	18.00	24.66	29.21
		1000	42.08	53.29	58.66	24.12	30.16	35.21
		2000	55.29	64.54	69.08	27.95	35.75	40.79
	200	500	25.16	40.12	48.21	17.45	25.75	30.54
		1000	40.33	54.54	58.45	23.83	30.50	34.83
		2000	55.04	65.37	70.83	28.00	35.25	40.37

- *Entries correspond to percentages across the number of replications and the set of test items.*

Table 4.7
Empirical Power Rates for Beaton's Fit Statistics (24 items)

Number of Replications (r)	Monte Carlo Samples (R)	Sample Size (N)	MR			MSR		
			α=1%	α=5%	α=10%	α=1%	α=5%	α=10%
100	100	500	26.21	38.95	46.50	12.79	21.12	25.21
		1000	40.91	54.04	60.91	19.91	23.79	28.62
		2000	56.16	64.71	70.45	22.21	26.91	32.08
	200	500	23.21	37.58	46.79	13.75	20.79	25.37
		1000	40.33	52.91	60.08	19.58	24.33	29.45
		2000	56.67	66.67	72.67	21.91	26.91	32.91
200	100	500	25.81	38.29	46.16	12.66	20.85	25.14
		1000	40.71	53.39	59.83	19.87	23.68	28.71
		2000	56.43	65.52	71.12	21.89	27.25	32.06
	200	500	24.79	39.06	47.49	11.50	18.56	23.06
		1000	40.62	53.37	60.33	19.77	24.29	29.18
		2000	56.83	67.04	73.06	22.02	27.58	33.45

- *Entries correspond to percentages across the number of replications and the set of test items.*

Table 4.8
Empirical Power Rates for Beaton's Fit Statistics (36 items)

Number of Replications (r)	Monte Carlo samples (R)	Sample Size (N)	MR			MSR		
			α=1%	α=5%	α=10%	α=1%	α=5%	α=10%
100	100	500	25.83	36.97	44.86	9.78	19.05	24.17
		1000	41.08	50.72	57.11	17.47	21.75	26.27
		2000	53.25	61.02	66.81	20.13	24.17	30.69
	200	500	24.33	39.02	46.50	9.75	19.08	23.61
		1000	38.55	50.02	56.25	17.80	21.86	26.69
		2000	53.55	62.81	68.17	19.61	24.42	30.25
200	100	500	25.90	37.61	45.40	9.87	19.15	23.83
		1000	40.47	50.55	57.25	17.40	21.72	26.31
		2000	53.12	61.73	67.58	20.06	24.73	31.09
	200	500	24.17	38.09	45.36	9.51	18.91	23.45
		1000	39.00	50.36	56.52	17.71	21.83	26.51
		2000	52.94	62.12	67.62	19.71	24.59	30.61

- *Entries correspond to percentages across the number of replications and the set of test items.*

As can be seen from Tables 4.6- 4.8, when H_0 was false for all the test items, Beaton's MR statistic detected more model misfit than Beaton's MSR statistic under the same test conditions. Beaton's MSR statistic detected some degree of model misfit, but had considerably less power than Beaton's MR statistic. For example, under the condition of 12 items test with 100 Monte Carlo resample size and sample size N equal to 500, the empirical power rate across 100 replications was ~ 38 for Beaton's MR statistic. Under the same condition, the empirical power rate for Beaton's MSR was only ~ 24. The empirical power based on Beaton's MR and MSR statistics increased as α increased from 0.01 to 0.10 and as sample size increased from 500 to 2000. Test length had a different effect on empirical power rates for Beaton's MR statistic than for Beaton's MSR statistic. For Beaton's MR statistic, there was little difference in empirical power rates for different test lengths. For Beaton's MSR statistic, the empirical power rates decreased as test length increased.

In summary, the empirical power rates were not very high for Beaton's MR and MSR statistics when H_0 was false for all the test items. The reason may be that there were little differences between the same slope parameters used to simulate item responses and the constant slope parameter used to calibrate and evaluate model-data-fit. For example, for 12 items test with sample size (N = 1000), the initial slope parameters for the six items used to simulate item responses were 0.49, 1.55, 1.10, 0.91, 2.08 and 0.89; while the constant slope parameter used to calibrating model and evaluating model-data-fit was 1.35.

Table 4.9 presents an example of empirical power rates for the 12 items test under the condition of 100 Monte Carlo samples and sample size N equal to 1000. Entries in the table represent the number of correct rejections across the number of replications for each item. As can be seen for items in which the slope parameter was similar to the estimated constant slope

parameter (1.35), power was low. However, for items in which the slope parameter differed from the estimated constant parameter, the observed power was high. For example, in Table 4.9, the correct rejections for the 3rd item were only 4 for Beaton's MR statistic under the conditions of a 12 item test, sample size N=1000, Monte Carlo sample size R = 100 and $\alpha = 0.05$. Under the same test conditions, the correct rejections for the 3rd item were 5 for Beaton's MSR statistic. In contrast, for the same conditions, the observed power was high for item 1.

Table 4.9
Empirical Power Rates for Beaton's Fit Statistics (test length=12; Monte Carlo samples =100, Sample size =1000)

Item #	Slope Parameter (α)	MR			MSR		
		α =1%	α = 5%	α =10%	α =1%	α =5%	α =10%
1	**.49**	63	83	91	99	100	100
2	**1.55**	34	53	63	0	0	0
3	**1.10**	1	4	5	1	5	7
4	**.91**	11	27	37	8	34	51
5	**2.08**	100	100	100	0	0	0
6	**.89**	88	97	99	0	0	0
7	**.49**	74	91	94	100	100	100
8	**1.55**	40	60	69	0	0	0
9	**1.10**	0	1	3	0	7	16
10	**.91**	10	21	28	10	29	41
11	**2.08**	81	95	97	0	0	0
12	**.89**	4	11	16	77	92	98
Average		42.16	53.58	58.5	24.58	30.58	34.41

- *Entries correspond to percentages of misfit across the number of replications and test items.*

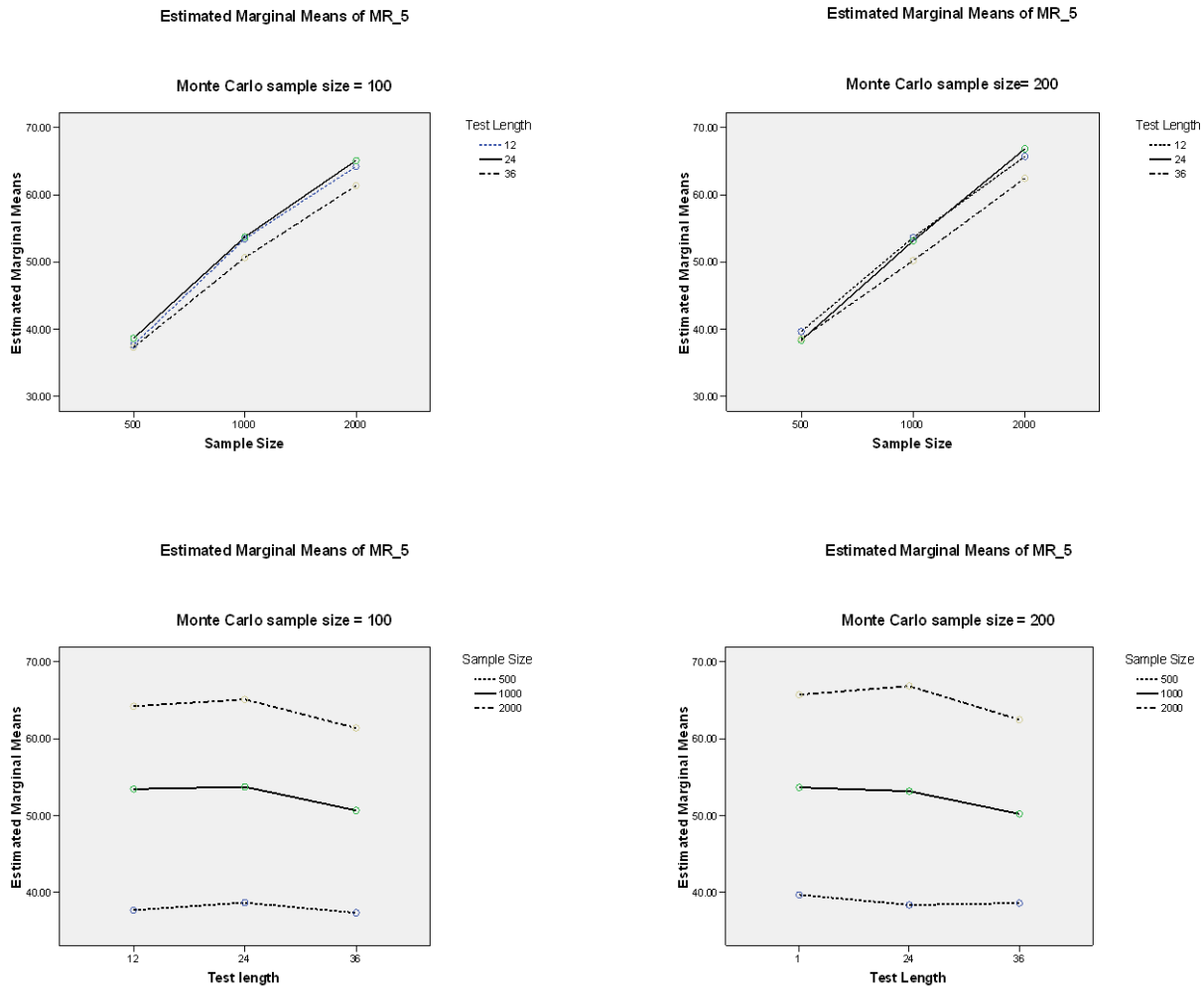
4.3.1.1 Analysis of Factor Effects

To examine the effect of different manipulated factors (e.g., test length, sample size, Monte Carlo sample size and number of replications) on empirical power, ANOVA tests for mean and two-way interaction effects were conducted. All higher-order interaction effects were excluded from the analysis since a preliminary analysis found there effects to have low effect sizes. The dependent variables were the empirical power rates at three α (0.01, 0.05 and 0.10) levels. The independent variables were the four manipulated factors. ANOVA test results for Beaton's MR and MSR statistics are provided in Tables 4.10-4.15. Note that main effect and significant interaction effects are reported in the ANOVA tables. Since the effect size indicates the relative importance of the main or interaction effect, effect sizes are also reported in the ANOVA test results. As an index of effect size, partial Eta squared was computed using SS_{effect} / (SS_{effect}+SS_{error}).

Table 4.10
ANOVA Test for the Empirical Power Rates Based on Beaton's MR Statistic (α=0.05)

Effect	Sum square	df	Mean Square	F	Sig.	Partial Eta Squared
Test Length	47.252	2	23.626	53.815	.000	.871
Monte Carlo Sample Size	4.673	1	4.673	10.644	.005	.399
Number of replications	.382	1	.382	.871	.365	.052
Sample Size	4050.575	2	2025.288	4613.145	.000	.998
Test Length × Sample Size	14.887	4	3.722	8.478	.001	.679
Monte Carlo Sample Size × Sample Size	4.726	2	2.363	5.382	.016	.402
Error	7.024	16	.439			

Table 4.10 presents the ANOVA results for Beaton's MR statistic at $\alpha = 0.05$ level. As can be seen, the interaction effects, test length x sample size and Monte Carlo sample size x sample size, were significant. Since the interaction effects were significant, the nature of interaction effects should be analyzed before determining the simple main effects. To analyze the nature of interaction effects, the cell means of empirical power rates for MR statistic were graphed in Figure 4.13. Figure 4.13 Mean Plots for MR Statistic with α=0.05



From Figure 4.13, similar patterns are observed for different Monte Carlo sample sizes (100 and 200). Results also indicated there was a main effect for sample size and there were no practical effects related to the Monte Carlo sample size and test length. For example, the average empirical power rates were 38.34, 52.46 and 64.69 for sample sizes of 500, 1000 and 2000, respectively. In contrast, the average empirical power rates were 53.00, 52.77 and 50.40 for test lengths of 12, 24 and 36.

Tables 4.11 and 4.12 present the ANOVA test results at α=0.01 and α=0.10, respectively. As can be seen results similar to when α =.05 were found. In addition, evaluation of interaction effects revealed similar interpretations. Thus for Beaton's MR statistic, sample size was the most important factor in determining the empirical power rates when H_0 was false for all the test items.

Table 4.11
ANOVA Test for the Empirical Power Rates Based on Beaton's MR Statistic (α=0.01)

Effect	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Test Length	14.552	2	7.276	35.110	.000	.814
Monte Carlo Sample Size	8.085	1	8.085	39.013	.000	.709
Number of replications	.011	1	.011	.052	.823	.003
Sample Size	5317.661	2	2658.830	12830.537	.000	.999
Test Length × Sample Size	12.478	4	3.119	15.053	.000	.790
Monte Carlo Sample Size × Sample Size	6.423	2	3.212	15.499	.000	.660
Error	3.316	16	.207			

Table 4.12
ANOVA Test for the Empirical Power Rates Based on Beaton's MR Statistic (α=0. 10)

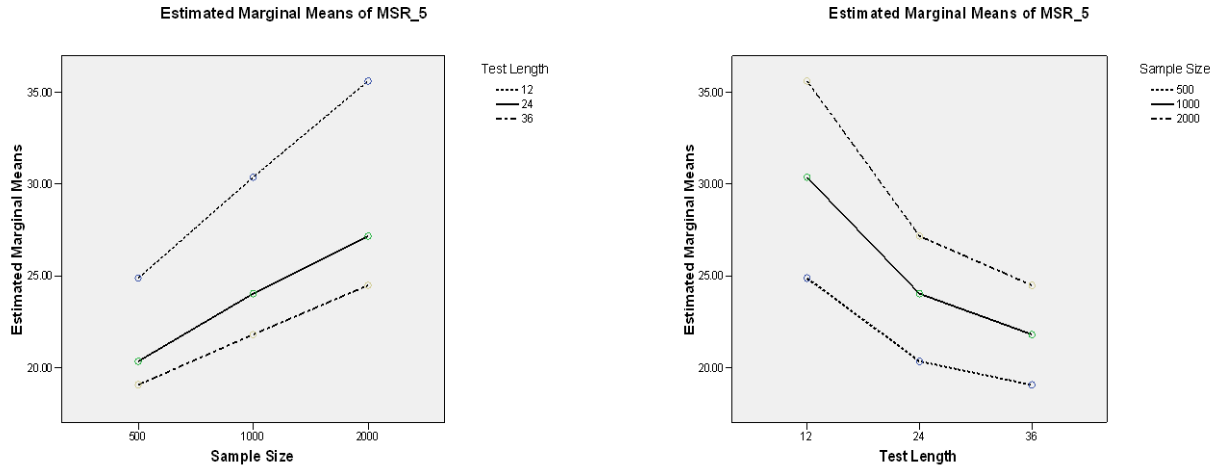
Effect	Sum Square	df	Mean Square	F	Sig.	Partial Eta Squared
Test Length	54.050	2	27.025	76.996	.000	.906
Number of replications	0.063	1	.063	.178	.679	.011
Monte Carlo Sample Size	7.093	1	7.093	20.209	.000	.558
Sample Size	3320.509	2	1660.255	4730.140	.000	.998
Test Length × Sample Size	10.969	4	2.742	7.813	.001	.661
Monte Carlo Sample Size × Sample Size	9.166	2	4.583	13.057	.000	.620
Error	5.616	16	.351			

Next, the factors' effect on empirical power rates for Beaton's MSR statistic was investigated. Table 4.13 presents the ANOVA test results for empirical power rates based on Beaton's MSR statistic at α=0.05. For the MSR statistic, the interaction effect of test length x sample size was significant. Because of the significant interaction effect, the cell means for empirical power rates based on Beaton's MSR statistic were plotted (see Figure 4.14).

Table 4.13
ANOVA Test for the Empirical Power Rates Based on Beaton's MSR Statistic (α=0. 05)

Effect	Sum Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Test Length	473.629	2	236.814	624.584	.000	.987
Number of replications	.004	1	.004	.009	.924	.001
Monte Carlo Sample Size	.007	1	.007	.018	.894	.001
Sample Size	353.291	2	176.646	465.893	.000	.983
Test Length × Sample Size	30.525	4	7.631	20.127	.000	.834
Error	6.066	16	.379			

Figure 4.14 Mean Plots for MSR Statistic with α=0.05



From figure 4.14, an ordinal interaction was observed since the nonparallel lines do not intersect with each other. Since an ordinal interaction was observed, a main effect can still be analyzed. Figure 4.14 indicated that both the sample size main effect and the test length main effect appeared significant. For example, the average empirical power rates for the three different test length were 30.33, 23.83 and 21.77 for 12, 24 and 36 items respectively. With regard to sample size, the average empirical power rates were 21.41, 25.39 and 29.08 for sample sizes of 500, 1000 and 2000, respectively.

Tables 4.14-4.15 present the ANOVA results of empirical power rates for Beaton's MSR statistic at α = 0.01 and α = 0.10, respectively. As can be seen from the tables, ANOVA test results obtained for α = 0.01 and α = 0. 10 were similar to α = 0.05. The same sources of effect were significant and the corresponding effect sizes were also close to each other for the three nominal α levels. In addition, the mean differences for each effect also displayed a similar pattern. Thus similar conclusions could be drawn regardless of which α level to use. As a result,

sample size and test length were two important factors in determining the empirical power rates for Beaton's MSR statistic when H_0 was false for all the test items.

Table 4.14
ANOVA Test for the Empirical Power Rates Based on Beaton's MSR Statistic (α=0. 01)

Source	Sum Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Test Length	340.808	2	170.404	920.902	.000	.991
Number of replications	.064	1	.064	.347	.564	.021
Monte Carlo Sample Size	.656	1	.656	3.546	.078	.181
Sample Size	631.641	2	315.821	1706.763	.000	.995
Test Length × Sample Size	4.304	4	1.076	5.815	.004	.592
Error	2.961	16	.185			

Table 4.15
ANOVA Test for the Empirical Power Rates Based on Beaton's MSR Statistic (α=0. 10)

Effect	Sum Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Test Length	427.280	2	213.640	533.195	.000	.985
Number of replications	0	1	0	.000	.994	.000
Monte Carlo Sample Size	.003	1	.003	.007	.936	.000
Sample Size	437.555	2	218.778	546.017	.000	.986
Test Length × Sample Size	17.872	4	4.468	11.151	.000	.736
Error	6.411	16	.401			

4.3.2 Empirical Power under the Condition that H_0 was False for a Subset of Test Items

Another type of model misfit was introduced when the null hypothesis was false for a subset of items, that is, a subset of item parameters of the calibrating model used to assess goodness-of-fit were altered from the item parameters used to simulate item responses. The consideration for this condition was that the presence of misfitting items may affect the evaluation of other items which truly fit the model. In this study, one slope and one threshold parameter of the calibrating model was altered from the item parameters used to simulate item responses. The parameters were changed for only one item to see the effect of the manipulated item on all the other non-manipulated items. The slope parameter of the first item was altered by 0.5. For a separate analysis, the first threshold parameter of item 11 was altered by 0.25.

Tables 4.16-4.18 summarize the rejection rates for Beaton's MR and MSR statistics under the conditions that the slope parameter of the calibrating model of the first item was altered by .5 from the item parameters used to simulate item responses. In Tables 4.16 – 4.18, since H_0 was false for item 1, entries for item 1 are the percentage of correct rejections across the number of replications, and therefore reflect an empirical power rate. Since H_0 was true for all the other non-manipulated items, the entries for all the other items are the average percentage of false rejections across the number of replications (Type I error rates).

Table 4.16
Rejection Rates for Beaton's Fit Statistics (altered item # =1, test length=12)

Number of Replication (r)	Monte Carlo sample (R)	Sample Size (N)	Misfit item #	MR			MSR		
				α=1%	α=5%	α=10%	α=1%	α=5%	α=10%
100	100	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining	11	21.81	32.91	1.45	5.36	11.63
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 11	22.09	37.09	47.91	1.72	6.45	12.09
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 11	34	52.91	64.18	1.81	5.45	10
	200	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 11	12.91	27.63	38.18	1.27	5.81	11.09
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 11	18.36	35.81	50.36	0.91	6	12.54
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 11	36	55.81	65.45	1.45	6.72	13.81
200	100	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 11	10	20.18	31	1.41	5.86	11.86
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 11	20.59	34.77	45.09	1.67	5.91	11.63
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 11	37.81	54.81	65.72	2.18	6.27	11
	200	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 11	10.72	25.36	35.09	0.91	5.91	10.54
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 11	18.36	34.45	46.63	0.91	6.54	12.54
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 11	37	57.54	66.27	1.54	6.72	13.09

Table 4.17
Rejection Rates for Beaton's Fit Statistics (altered item # =1, test length=24)

Number of Replications (r)	Monte Carlo sample (R)	Sample Size (N)	Misfit item #	MR			MSR		
				α=1%	α=5%	α=10%	α=1%	α=5%	α=10%
100	100	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	5.13	11.73	17.56	1.47	6.52	11.82
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	6.52	13.95	22.39	1.04	4.73	9.69
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	10.32	22.59	32.76	1.01	5.27	9.47
	200	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	2.69	9.39	17.04	0.86	4.26	9.30
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	5.30	12.69	21.39	0.78	4.17	9.04
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	10.13	23.91	38.26	1.39	4.34	9.82
200	100	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	4	9.34	15.82	0.95	4.65	9.13
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	7.59	15.77	24.09	1.68	5.91	11.63
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	10.95	22.21	31.17	1.43	5	10.04
	200	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	3.04	10.47	18.52	0.69	4.34	9.39
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	4.73	12.08	20.17	0.91	3.86	8.73
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 23	9.17	25.47	36.47	0.95	3.95	9.26

Table 4.18
Rejection Rates for Beaton's Fit Statistics (altered item # =1, test length=36)

Number of Replications (r)	Monte Carlo sample (R)	Sample Size (N)	Misfit item #	MR			MSR		
				α=1%	α =5%	α =10%	α =1%	α =5%	α =10%
100	100	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	1.71	6.57	13.02	0.74	4.17	8.57
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	3.6	8.97	14.34	1.25	4.85	9.65
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	6.45	14.11	21.54	0.74	3.94	8.45
	200	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	1.94	7.88	13.82	0.97	4.34	8.8
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	2.34	8.4	14.85	0.91	4.28	9.42
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	4.4	12.11	19.08	1.2	5.14	9.42
200	100	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	1.71	6.14	11.65	0.68	4.17	9.22
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	3.74	9.57	16.2	1	4.42	9.11
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	5.48	12.68	20.05	0.77	4.51	8.71
	200	500	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	1.6	7.11	12.8	1.05	4.45	8.71
		1000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	2.37	8.45	15.05	0.94	4.4	9.11
		2000	Item 1 (a+.5)	100	100	100	100	100	100
			Remaining 35	3.74	11.25	18.54	0.6	3.31	7.48

- *Entries correspond to percentages across the number of replications. For unaltered items, percentages represent an average across the set of unaltered items.*

As can be seen in Tables 4.16 - 4.18, for the manipulated item with an altered slope parameter (.49 to .99), high empirical power rates were obtained for Beaton's MR and MSR statistics since the percentage of correct rejections were all 100 across the number of replications. For all the other non-manipulated items, different results were obtained for Beaton's MR and MSR statistics. For Beaton's MR statistic, the rejection rates for all the non-manipulated items were larger than the corresponding expected nominal rejection rates (α). In contrast, for Beaton's MSR statistic, approximate nominal rejection rates were observed for all the non-manipulated items. The results indicated that when one item was purposely manipulated as misfitting, the misfitting item may have had some impact on the non-manipulated items for Beaton's MR statistic, but had no impact on the non-manipulated items for Beaton's MSR statistic. For example under the combination of 12 items test, sample size N=500 and Monte Carlo sample size R=100 with $\alpha = 0.05$, the average percentages of false rejections across all the non-manipulated items were 21.81 for Beaton's MR statistic and 5.36 for Beaton's MSR statistic. Since approximate nominal type I error rates for all the non-manipulated items were observed for Beaton's MSR statistic, there was little difference due to different factors (e.g. test length, sample size, Monte Carlo sample size and number of replications). However, the rejection rates of Beaton's MR statistic for the non-manipulated items were affected by test length and sample size. For Beaton's MR statistic, the rejection rates for all the non-manipulated items decreased as test length increased. For example with $\alpha=0.05$, the rejection rates for Beaton's MR statistic, with sample size (N = 500) and Monte Carlo sample size (R = 100) for 100 replications, were 21.81 for 12 items test and 6.57 for 36 items test. Sample size was another factor which affected the rejection rates of all the non-manipulated items for Beaton's MR

statistic. As can be found from Tables 4.16-4.18, the rejection rates for all the non-manipulated items for Beaton's MR statistic increased as sample size N increased from 500 to 2000. However, there were little differences in empirical power rates for different Monte Carlo sample sizes and number of replications.

The finding of more approximate nominal Type I error rates for the MR statistic as test length increased may indicate that the precision of parameter estimates is playing a role. In this study, when one item was manipulated as misfitting, more influence of the manipulated item on the parameter estimates for non-manipulated items may have occurred in shorter tests than longer tests. While the manipulated item appeared to affect the non-manipulated items for Beaton's MR statistic (a signed statistic), but this same effect was negligible for Beaton's MSR statistic (an unsigned statistic).

In summary, when the item slope parameter was altered, high empirical power for the manipulated item and approximate nominal rejections rates for all the other non-manipulated items were observed for Beaton's MSR statistic. While high empirical power for the manipulated item was observed for the MR statistic, inflated type I error rates for the non-manipulated items were observed especially for shorter tests. Therefore, Beaton's MSR performed better than Beaton's MR statistic regardless of test length and sample size when slope parameter of the calibrating model was altered by .5. It should be noted that in this study only an item with low slope parameter was altered. The results could be different if a different slope parameter was altered.

The results for altering the threshold parameter for a test item were also investigated. Tables 4.19-4.21 present rejection rates for Beaton's MR and MSR statistics under the condition that an item threshold parameter for the calibrating model was altered from the item parameters

used to simulate item responses. In this case, the first threshold parameter for item 11 was altered (0.33 to 0.58). Since H_0 was false for item 11, the entries for item 11 reflect empirical power rates. The entries for all the other non-manipulated items are the average percentage of false rejections across the number of replications (Type I error rates) since H_0 was true for all the other test items.

Table 4.19
Rejection Rates for Beaton's Fit Statistics (altered item # =11, test length=12)

Number of Replications (r)	Monte Carlo sample (R)	Sample Size (N)	Misfit Item #	MR			MSR		
				α=1%	α=5%	α=10%	α=1%	α=5%	α=10%
100	100	500	Item 11 (b_2+.25)	99	99	99	8	29	41
			Remaining 11	38.09	55.18	64.54	1.63	8.36	15.63
		1000	Item 11 (b_2+.25)	100	100	100	15	49	67
			Remaining 11	58.27	75	82	2.63	7.63	14.18
		2000	Item 11 (b_2+.25)	100	100	100	27	57	69
			Remaining 11	84.45	94.36	96.81	3.36	11.54	18.09
	200	500	Item 11 (b_2+.25)	99	100	100	6	27	41
			Remaining 11	30.18	53.45	63.36	1	7.45	14.54
		1000	Item 11 (b_2+.25)	100	100	100	17	46	70
			Remaining 11	55.36	74.45	82.18	2	8	14.09
		2000	Item 11 (b_2+.25)	100	100	100	30	68	84
			Remaining 11	86.36	94.54	96.54	3.27	12.09	18.72
200	100	500	Item 11 (b_2+.25)	98.5	99	99	7.5	29	44
			Remaining 11	33.18	50.91	61.04	1.68	7.68	14
		1000	Item 11 (b_2+.25)	100	100	100	16	48.5	64.5
			Remaining 11	57	74.31	81.04	2.27	8.04	14.77
		2000	Item 11 (b_2+.25)	100	100	100	29	67.5	78.5
			Remaining 11	84.36	93.5	96.27	3.41	11.45	18.5
	200	500	Item 11 (b_2+.25)	98.5	99.5	99.5	8	28	46.5
			Remaining 11	31.54	53.86	64.13	1.04	7.09	14.27
		1000	Item 11 (b_2+.25)	100	100	100	14	45.5	67
			Remaining 11	58.45	77.95	85.31	1.81	8.63	15.59
		2000	Item 11 (b_2+.25)	100	100	100	31.5	66.5	82.5
			Remaining 11	85.95	93.91	96.18	3.72	13.18	20.36

Table 4.20
Rejection Rates for Beaton's Fit Statistics (altered item # =11, test length=24)

Number of Replications (r)	Monte Carlo sample (R)	Sample Size (N)	Misfit item #	MR			MSR		
				α=1%	α=5%	α=10%	α=1%	α=5%	α=10%
100	100	500	Item 11 (b_2+.25)	100	100	100	7	32	49
			Remaining 23	12.30	26.86	38.04	1	5.61	11.21
		1000	Item 11 (b_2+.25)	100	100	100	14	37	62
			Remaining 23	21.30	38.78	50	1.13	6.52	13.30
		2000	Item 11 (b_2+.25)	100	100	100	19	56	77
			Remaining 23	51.21	68.17	77.13	3.21	7.65	14
	200	500	Item 11 (b_2+.25)	100	100	100	5	28	47
			Remaining 23	9.21	25.73	38.69	0.69	5.04	10.95
		1000	Item 11 (b_2+.25)	100	100	100	11	38	58
			Remaining 23	19.86	40.52	53.78	1.17	6.08	11.69
		2000	Item 11 (b_2+.25)	100	100	100	27	66	78
			Remaining 23	41.78	63	73.04	1.52	7.21	13.43
200	100	500	Item 11 (b_2+.25)	100	100	100	6.5	30	47.5
			Remaining 23	11.06	24.5	35.11	1.06	5.24	10.61
		1000	Item 11 (b_2+.25)	100	100	100	12	35	58
			Remaining 23	23	41.52	53.43	1.56	6.43	12.69
		2000	Item 11 (b_2+.25)	100	100	100	22	61	79
			Remaining 23	47.04	64.78	74.26	1.56	7.73	13.95
	200	500	Item 11 (b_2+.25)	100	100	100	12.5	29	52
			Remaining 23	9.13	24.26	35.21	0.82	5.13	11.39
		1000	Item 11 (b_2+.25)	100	100	100	9	35.5	55
			Remaining 23	20.41	41	53.68	1.22	6.15	11.29
		2000	Item 11 (b_2+.25)	100	100	100	24.50	61	77
			Remaining 23	43.83	64.62	74.78	1.55	7.12	13.48

Table 4.21
Rejection Rates for Beaton's Fit Statistics (altered item # =11, test length=36)

Number of Replications (r)	Monte Carlo sample (R)	Sample Size (N)	Misfit item #	MR			MSR		
				α=1%	α=5%	α=10%	α =1%	α=5%	α=10%
100	100	500	Item 11 (b_2+.25)	100	100	100	6	24	42
			Remaining35	5.2	14.85	24.45	1.25	5.54	10.74
		1000	Item 11 (b_2+.25)	100	100	100	24	46	72
			Remaining35	11.37	24.28	34.45	0.8	5.37	12.17
		2000	Item 11 (b_2+.25)	100	100	100	29	69	82
			Remaining35	25.16	45.37	55.28	1.25	7.25	11.85
	200	500	Item 11 (b_2+.25)	100	100	100	9	20	52
			Remaining35	3.54	14	21.77	0.85	5.42	11.2
		1000	Item 11 (b_2+.25)	100	100	100	21	56	70
			Remaining35	12.91	29.82	42.97	0.62	4.62	11.6
		2000	Item 11 (b_2+.25)	100	100	100	26	74	88
			Remaining35	26.62	48.4	59.93	1.31	6.62	12.74
200	100	500	Item 11 (b_2+.25)	99.5	100	100	10	30	43
			Remaining35	6.17	16.14	25.51	1.02	5.02	10.71
		1000	Item 11 (b_2+.25)	100	100	100	21	45	68
			Remaining35	13.34	27.82	38.17	1.11	5.48	12.08
		2000	Item 11 (b_2+.25)	100	100	100	36	74	88
			Remaining35	29.22	46.74	58.48	1.74	6.65	13.05
	200	500	Item 11 (b_2+.25)	100	100	100	7	26	47
			Remaining35	4.28	15.25	24.2	0.74	5.22	11.25
		1000	Item 11 (b_2+.25)	100	100	100	18	54	69
			Remaining35	12.91	29.88	42.05	0.65	5.14	11.94
		2000	Item 11 (b_2+.25)	100	100	100	28	71	92
			Remaining35	28.45	50.05	60.4	1.6	7.02	14.11

- *Entries correspond to percentages across 100 replications. For unaltered items, percentages represent an average across the set of unaltered items.*

As can be seen from Tables 4.19-4.21, the results based on Beaton's MR statistic were different from the results based on Beaton's MSR statistic for both the manipulated and non-manipulated items under the condition of altering the first threshold parameter of item 11 by .25. For the manipulated item 11, Beaton's MR statistic displayed considerably more power than Beaton's MSR statistic. The empirical power rates for item 11 based on Beaton's MR statistic were all close to 100 across all the experimental conditions. In contrast, for Beaton's MSR statistic, the empirical power rates for item 11 were not high across the experimental conditions. With regard to empirical power based on MSR statistic for item 11, results indicated an increase in empirical power as α went from 0.01 to 0.10, and as test length increased from 12 to 36 and sample size increased from 500 to 2000, but there was no differences in empirical power across different Monte Carlo resample sizes and number of replications.

For all the other non-manipulated items, approximate nominal rejection rates were observed for Beaton's MSR statistic, which was as expected. However, as found for the MR statistic, large rejection rates were observed for Beaton's MR statistic. For example when α=0.05, the rejection rates was 55.18 for Beaton's MR statistic for 12 items test with sample size (N=500) and Monte Carlo sample size (R=100) across100 replications. While under the same condition, the rejection rates were only 8.36 for Beaton's MSR statistic. Therefore, for Beaton's MSR statistic, the misfitting item appeared to have little impact on the non-manipulated items and nominal Type I error rates were observed for all the non-manipulated items regardless of different factors (e.g., test length, sample size, Monte Carlo sample size and number of replications). However, for Beaton's MR statistic, the manipulated item may have affected the non-manipulated items, especially for short test. For the rejection rates based on Beaton's MR

statistic for all the non-manipulated items, there was little difference in rejection rates for different Monte Carlo sample sizes and number of replications. However, the false rejection rates increased as nominal α increased from 0.01 to 0.10 and as sample size N went from 500 to 2000, and decreased rapidly as test length increased. For example when α=0.05, the rejection rates were 55.18 for the 12 item test and 14.85 for the 36 item test with sample size (N=500) and Monte Carlo sample size (R=100) across100 replications.

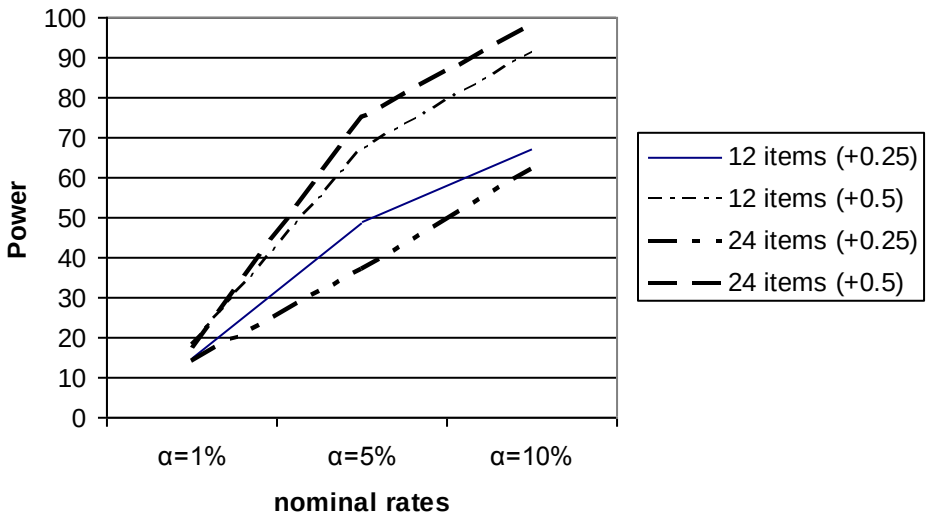
From the comparison of the results of Beaton's fit statistics for the manipulated item (item 11), it may be that Beaton's MR statistic over amplified the misfitting effect and Beaton's MSR statistic generated more reasonable empirical power rates than Beaton's MR statistic. Changing threshold parameter by 0.25 may not reflect such a pronounced misfit condition in the context of polytomous models. Therefore, the case for altering the threshold parameter by .5 was investigated. Table 4.22 and Figure 4.15 present the comparison of altering first threshold parameter of item 11 by .25 with .5 under the condition with number of replications=100, Monte Carlo sample size=100 and sample size =1000.

Table 4.22
Results for Altering First Threshold Parameter of Item 11 by .25 with .5

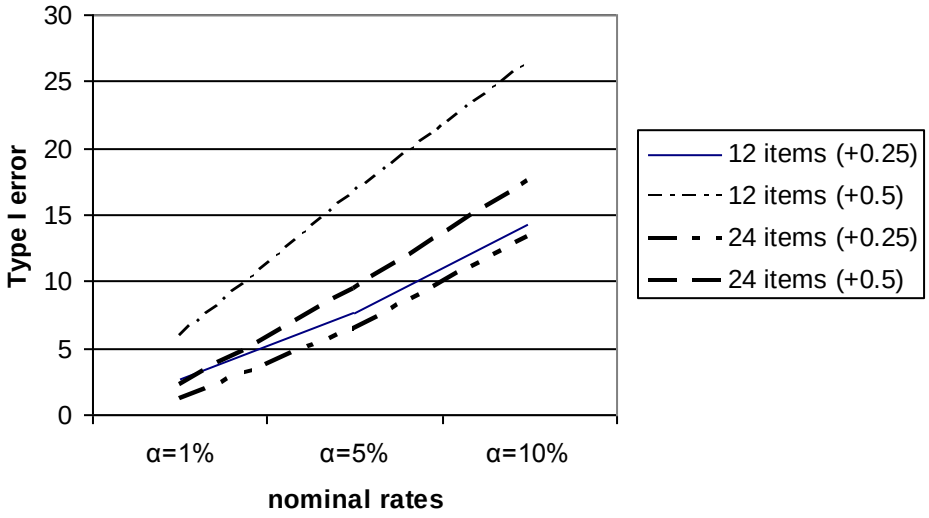
Test length	**Misfit item #**	**MR**			**MSR**		
		α=1%	**α=5%**	**α=10%**	**α=1%**	**α=5%**	**α=10%**
12	**Item 11 (b_2+.25)**	100	100	100	15	49	67
	Remaining 11	58.27	75	82	2.63	7.63	14.18
	Item 11 (b_2+.5)	100	100	100	18	67	91
	Remaining 11	97.09	99.27	99.72	5.91	16.72	26.45
24	**Item 11 (b_2+.25)**	100	100	100	14	37	62
	Remaining 23	21.30	38.78	50	1.13	6.52	13.30
	Item 11 (b_2+.5)	100	100	100	17	75	98
	Remaining 23	69.47	84.78	90.95	2.26	9.47	17.56

Figure 4.15
Rejection rates for MSR statistic for altering first threshold parameter of item 11 by .25 with .5

a) Manipulated item (item 11)



b) Non-Manipulated items



From Table 4.20 and Figure 4.15, higher empirical power rates for MSR statistic were observed when altering the first threshold parameter of item 11 by .5 than by .25, which is just as expected. Although there was some inflated effect on all the non-manipulated items, the inflated effect tended to decrease as test length increased. Thus, the Beaton MSR statistic also provided more reasonable overall results under the conditions of altering the threshold parameter. As a summary, Beaton's MSR statistic was better than the MR statistic in detecting model-misfit. One reason may be related to the sampling distributions for these statistics. Recall that the MSR was more variable than the MR statistic (see Tables 4.1 and 4.2). Therefore, the MSR statistic may have been less sensitive to small differences between the fit statistic and the critical values derived from the sampling distributions.

CHAPTER 5

SUMMARY AND DISCUSSION

5.1. Purpose and Findings

With IRT being widely used in educational and psychological testing, the evaluation of IRT goodness-of-fit is really important for validating the use of IRT models. Since the misfit between an IRT model and empirical data may potentially threaten the realization of IRT model advantages, it is important that model-data-fit be evaluated before model applications.

Traditional IRT goodness-of-fit methods use a chi-square test to evaluate the difference between observed and expected score distributions. The problems associated with these methods are the imprecise ability estimation and sometimes poor approximation of the null distribution. To avoid this problem, Beaton (2003) proposed two fit statistics (MR and MSR). Beaton's fit statistics are based on a standardized residual calculated from an expected and observed response. Different from tradition methods which use a null distribution to test the hypothesis, Beaton proposed using resampling method to generate the sampling distribution and test the hypothesis.

Beaton's new fit statistics have some advantages: (1) Avoid the arbitrariness of subdividing the ability scale into interval and classifying of examinees into ability subgroups; (2) Use five plausible ability values to account for imprecision of ability estimation; (3) Evaluate the goodness-of-fit for both dichotomous and polytomous items; (4) Use a Monte Carlo resampling

method to generate the sampling distribution for the fit statistic and test the hypothesis. This avoids any inappropriately defined null distribution and critical values.

There has been some previous research evaluating model-data-fit for NAEP assessment using Beaton's fit statistics. Dresher (2004) evaluated the model-data-fit for a sample of students from 2003 NAEP mathematics assessment for grade 4 and 8 using MR statistic. The results indicated that nominal rejection rates were obtained when a sample was randomly drawn from the nation. Li's (2006) dissertation examined model-data-fit for students with Limited English proficient (LEP) and students with disabilities (SD) in NAEP 2000 grade 8 mathematics state assessments. Although Beaton's fit statistics have been applied to evaluate model-data-fit, no research has been conducted to evaluate their performance under varied testing conditions. Therefore, the objective of this study was to evaluate the performance of Beaton's MR and MSR statistics. To accomplish this goal, a Monte Carlo simulation study was implemented to investigate the sampling distribution, Type I error rates and empirical power for Beaton's fit statistics.

The results of this study indicated that the sampling distributions of the test items for Beaton's fit statistics belonged to the family of normal distributions. However, there was no basis for a theoretical sampling distribution to test the hypothesis of model-data-fit. Therefore, a Monte Carlo resampling method would be required to test the hypothesis of model-data-fit for Beaton's fit statistics.

Results of this study also indicated that Type I error rates for Beaton's fit statistics were close to the nominal rate α when using the resampling-based method for hypothesis testing. The approximate nominal Type I error rates were not affected by different test lengths (12, 24 or 36

items), sample sizes (500, 1000 or 2000), Monte Carlo resampling sizes (100 or 200) or number of replications (100 or 200).

Empirical power was investigated when the null hypothesis was false for all items and when it was false for one item only. For the empirical power rates under the condition that H_0 was false for all the items, higher power was observed for Beaton's MR statistic than Beaton's MSR statistic. Empirical power was not affected by different Monte Carlo resample sizes (100 or 200) and different number of replications (100 or 200). But empirical power increased as sample size increased. Also the effect of test length on empirical power rates for Beaton's MR statistic was different from Beaton's MSR statistic. For the MR statistic, there was little difference in empirical power for the different test lengths, whereas for the MSR statistic, empirical power decreased as test length increased.

For the empirical power rates under the condition that H_0 was false for one item, adequate power for the manipulated item was observed for Beaton's MR and MSR statistics under the condition of altering the slope parameter by .5. However, when altering the threshold parameter by .25, only the MSR statistic showed more reasonable power to detect the model misfit. In addition, for all the non-manipulated items, more false rejections than expected were obtained for Beaton's MR statistic, which implied the manipulated item may have some effect on altering model-data-fit for the non-manipulated items. This effect was also more serious under the condition of altering the threshold parameter by 0.25 than altering the slope parameter by 0.5. Finally, nominal rejection rates were observed for the MSR statistic as expected. The manipulated item has little or no effect on the examination of model-data-fit for the non-manipulated items using Beaton's MSR statistic.

5.2. Recommendations for Applied Researchers

Beaton's MSR statistic appears to offer an alternative to some previous methods for assessing goodness-of-fit. For example, when comparing results for Beaton's MSR statistics in this study with results for the rescaling method based on posterior ability distribution and Orlando and Thissen's method (Stone & Zhang, 2003), similar nominal Type I error rates and moderate to high empirical power were observed. In addition, Beaton's fit statistics are easy to compute and can be applied to assess goodness-of-fit for both dichotomous and polytomous IRT models. The most promising feature of Beaton's MSR statistic is that it can be used to assess goodness-of-fit for both shorter (12 items) and longer test (36 items). Based on the results of this study, the recommended sample size to assess model-data-fit is 500 or more, and a Monte Carlo resample size of 100 should be adequate for hypothesis testing.

Although nominal Type I error rates and high empirical power rates were obtained for Beaton's MR statistic, it would only be recommended for assessing goodness-of-fit for longer test (>36 items). For shorter test, the results of this study indicated that non-fitting items may affect the evaluation of model-data-fit. Similar to Beaton's MSR statistic, the recommended sample size to assess model-data-fit would be 500 or more with a Monte Carlo resample size of 100.

5.3. Limitations

Although Beaton's fit statistics have some advantages, the major limitation is that the resampling procedure is computer intensive. Another limitation is related to the interpretation of the results. Since this study was conducted based on specific simulation conditions and used a specific set of item parameters, the results may not generalize to other testing situations.

5.4. Suggestions for Future Research

This research investigated the performance of Beaton's fit statistics using a Monte Carlo resampling method. There are several possible directions for future research. One direction is related to the resampling method. In this study, the empirical power was not very high when H_0 was false for all the test items. Perhaps a different resampling method could be used to obtain higher power. For example, Dresher (2004) used a jackknife procedure to evaluate the model-data-fit for NAEP assessment using Beaton MR statistic.

The second direction is related to the plausible abilities used in Beaton's fit statistics. In this study, five plausible abilities as Beaton proposed were used to account for the imprecision of ability estimate. Further research could be conduct to investigate the optimal number of plausible abilities to achieve better statistical properties for Beaton's fit statistics.

The third direction is related to the empirical power when H_0 was false for a subset of test items. In this study, a slope difference of .5 for a dichotomous item and a threshold difference of .25 for a polytomous item were manipulated. For the .5 slope difference for the dichotomous item, similar power was obtained for Beaton's MR and MSR statistics. However, for the .25 difference of the threshold parameter for the polytomous item, different power was obtained for Beaton's MR and MSR statistics. To see if the difference in power is due to the different IRT model or altered different item parameter, a comparison of the power could be conducted for a dichotomous vs. a polytomous item under the condition of altering the same parameter with same difference.

REFERENCES

Andries,L. van der Ark(2001). Relationships and Properties of Polytomous ItemResponse Theory Models. *Applied Psychological Measurement*, vol.25, 273-282.

Baker, Frank (2001). The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD. http://edres.org/irt/baker/

Beaton, A.E. (2003). A procedure for testing the fit of IRT models for special populations. Unpublished manuscript.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37(1), 29-51.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22 (3), 265-289.

Dresher, A.R., Brenda S.-H. Tay-lim, Thind, S.K., Tang, Y. & Beaton, A.E.(2004) The Beaton Fit Index for Special Populations: An Application in NAEP. (Available from Educational Testing Service)

Embretson, S.E. & Reise, S.P. (2000). Item response theory for psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 147-200). New York NY: American Council on Education & Macmillan Publishing.

Harwell, M., Stone, C.A., Hsu, T.C. & Kirisci, L. (1996). Monte Carlo Studies in Item Response Theory, Applied Psychological Measurement, 20 (2), 101-125.

Hambleton, R. K., Swaminathan, H., & Rogers, Jane H. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage Publications, Inc.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. Journal of the American Statistical Association, 73, 253–263.

Li, Jie (2005). Examining the Effect of Accommodations For Special-needs Students in NAEP Through Model-data fit Analysis. (Dissertation)

Linden, W.J. & Hambleton, R. K (1996). Handbook of Modern Item Repsons Theory. New York: Springer.

McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 19, 49-57.

Orlando, Maria, PhD (1997). Item fit in the context of item response theory. Dissertation.

Orlando, M. & Thissen, D (2000). Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement, 24, 50-64.

Ostini, R. & Nering L. M. (2006). Polytomous item response theory models. London: SAGE PUBLICATIONS.

Rogers, J. (1994). RESID. Assessment of Fit for Unidimensional IRT Models. Program developed at the University of Massachusetts School of Education.

Samejima F. (1997). Graded Response model. Handbook of modern item response theory (pp.85-100).New York. Springer.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 79-98.

Stone, C.A., Mislevy, R.J., & Mazzeo, J. (April, 1994). Classification error and goodness-of-fit in IR models. Paper presented at the annual meeting of the American Educational Research Association, April, New Orleans.

Stone, C.A & Hansen, M.A. (2000a). The effect of errors in estimating ability on goodness of fit tests for IRT models. Educational and Psychological Measurement, 60, 974-991.

Stone, C.A.(2000b). Monte-carlo based null distribution for an alternative fit statistic. Journal of Educational Measurement, 37, 58-75.

Stone, C.A. (2000c). Empirical power and Type I error rates for a goodness-of-fit statistic based on posterior expectations and resampling-based inference. Educational and Psychological Measurement, 63, 566-583.

Stone,C.A. (2003).Empirical power and Type I error rates for an IRT fit statistic that considers the precision of ability estimates. Educational and Psychological measurement. Vol. 63(4), 566-583.

Stone, C.A. & Zhang, B. (2003). Comparing three new approaches for assessing goodness-of-fit in IRT models. Journal of Educational Measurement, 4, 331-352.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5(2), 245-262.

Yen, W.M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. Applied Psychological Measurement, Vol. 8(2), 125-145.