FIT INDICES FOR THE RASCH MODEL

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of the Ohio State University

By

Judit Antal, M.A.

* * * * *

The Ohio State University
2003

Dissertation Committee:

Professor Ayres G. D'Costa, Adviser

Professor William E. Loadman

Professor In Jae Myung

Approved by

Adviser
College of Education

ABSTRACT

This dissertation introduces a new family of non-parametric fit tests for the Rasch model combining the elements of Monte Carlo method, traditional hypothesis testing, and Item Response Theory model fit. The new tests do not make assumptions regarding the distributions of their test statistics. Rather, distributions used for testing model fit are generated "on the fly" using a Monte Carlo method. The developmental phases and algorithms for performing the tests are discussed in detail. Differences between the new method and the usually accepted residual-based fit tests are presented from theoretical and practical perspectives, as well as the correspondence between the new indices and the general case of fit analysis. Comprehensive validity and stability studies are conducted using real and computer simulated test data to demonstrate the performance of the proposed indices under various conditions and to make comparisons with previously used Rasch fit indices. The results of the new global fit analysis, also introduced in this thesis, show that when fit analysis is performed with the aid of the new tests the Rasch model performs very well. It is demonstrated using several test scenarios that the traditional mean-square fit index reports false misfit quite frequently. Although, the Monte Carlo p-values are always approximate, a stability study conducted in this dissertation reveals, that they show remarkable

stability with respect to the number of simulated matrices. It is shown, that for moderately sized response matrices a satisfactorily stable p-value can be obtained within a reasonable computing time, making the newly proposed technique available to the test developing community.

Édesanyámnak és Édesapámnak

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my adviser, Professor Ayres G. D'Costa, for his support throughout my study. His patience and encouragement helped me complete this work.

I also wish to thank my committee members, Professors In Jae Myung and William E. Loadman, for their valuable comments and professional insight.

I am grateful to Professor Andrea Kárpáti, a distinguished researcher and excellent teacher at the Loránd Eötvös University in Hungary, who introduced me to the field of educational testing and research.

I also wish to express my gratitude to my parents, István Kovács and Klára Füsi, who always supported my professional plans by giving me their trust and encouragement.

Finally, my special thanks go to my husband, Tamás, and my sweet little daughter, Panni, for their love, smile and endless support throughout these years.

VITA

December 4, 1971	Born - Győr, Hungary
1997	B.A. Education, Loránd Eötvös University, Budapest, Hungary
1999 - 2000	Graduate Research Associate, Quantitative Research, Evaluation and Measurement, College of Education, The Ohio State University
2000	M.A. Education, Quantitative Research, Evaluation and Measurement, College of Education, The Ohio State University
2000 - recent	Graduate Administrative Associate, Management Information Analysis and Reporting, The Ohio State University

PUBLICATIONS

Antal, J. (1998). Comparative educational entries. In I. Falus (Ed.), *New Hungarian Lexicon of Pedagogy.* Budapest, Hungary: Keraban.

Kárpáti, A., & Kovács, J. (1995). The Visual Narrative Test as an expressive means for problems and anxieties of children. In O. Scholz (Ed.), *Anxiety and Fear in Children's Art Works - Angst und Schrecken in der Kunsterziehung.* Berlin, Germany: Hochschule der Künste.

FIELDS OF STUDY

Major field: Education

Specialization: Educational Measurement and Statistics

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

A major purpose of measurement in the social sciences is to establish a linear ordering among objects such as persons (often referred to as subjects), examination questions (often referred to as items) and performances. There is a large collection of methods developed in the behavioral sciences to define and measure this wide variety of variables. The problem is twofold. On the one hand, *a priori* one does not know what to measure. Entities in human measurement do not come in an easily manageable format. It is a customary point of view, that a variable is not only measured in the process, but the measurement process itself is the definition of the measured variable. That is, say, an intelligence test only measures the intelligence of a person that is defined by that very test. This situation makes the task challenging but it also opens the door for a very exciting world of measurement theories. Of course, the final goal of research in the behavioral sciences is to furnish more universal tests that can be repeated over time. It is desirable to have test results that can be compared over a chain of re-takes. Test theories are concerned with the mathematical foundation of this experimental process (Allen & Yen, 1979; Hambleton & Swaminathan, 1985; Baker, 1992; Embretson & Hershberger, 1999; Rust, 1999).

The test theory of choice in this dissertation is item response theory (IRT) (Wright & Stone, 1979; Wright & Masters, 1982; Baker, 1982; Hambleton & Swaminathan, 1985; van der Linden, 1997). The seminal idea of IRT is that the extremely complex interaction between the subject and the item to be answered is captured by a set of *response probabilities* - usually presented in a form of the *response probability matrix* - governing the possible outcomes (*response matrices*). That is, the model does not make any attempt to explain why the subject gave a certain answer. It only tries to assign a probability to each outcome. Put in yet another way, in IRT one does not assume the existence of the true score or error score. It is not assumed that the response is uniquely determined by the subject-item pair. Even under idealistic test-taking situations the model carries a certain degree of indeterminacy considering the test outcome. This gives the model a great flexibility and was, and continues to be the key to its success.

IRT models are most easily classified by their item characteristic functions (ICF). In short, ICF gives the probability of a person giving a specific answer to an item. In general, it can depend on any number of parameters of the subject and/or the item (Baker, 1992).

Many IRT models assume that persons have only one variable affecting the ICF. In these cases this random variable is called *ability* and the model is called *unidimensional.*

Frequently, these IRT models also assume that the more ability subjects have, the more likely that they are to successfully complete a task requiring that ability. This is called the *monotonicity of ability,* and is expressed by the fact that the ICF is a

strictly increasing function of ability. It would be extremely counter-intuitive to work with a model which lacks monotonicity.

A family of item response models have been developed out of the work of a Danish mathematician George Rasch who first published his revolutionary ideas in 1960 (Rasch, 1960). His work was inspired by L. L. Thurstone (1925) and E. L. Thorndike (1926) who basically worked with similar measurement models and described the fundamental logic of psychological measurement. The main feature of the Rasch model is that its ICF depends only on a single item parameter, called *item difficulty.* In addition to the monotonicity of ability it features the similarly defined *monotonicity of difficulty.* This is rarely true in other IRT models. It is also counter-intuitive to drop monotonicity of difficulty. This topic is still a major source of hot debate over the possible IRT models.

The Rasch model is considered to be the first member of a big family of logistic models (Rasch, 1960; Wright & Stone, 1979; Wright & Masters, 1982). The family consists of the Rasch model (a.k.a. 1PLM or 1 parameter logistic model) the 2PLM, the 3PLM, and the 4PLM. They are named after the number of item parameters used in the ICF.

The theory of model fit in statistics has been opened up long time ago and been investigated from several perspectives (Drasgow & Levin, 1985; 1986; Drasgow & Levin & McLaughlin, 1987; Baker, 1992; Meijer & Sijtsma, 2002). Theory of fit in psychometric research has become a popular field in the 1960's and 1970's and has led to the discovery of several new fit statistics, or appropriateness indices. As

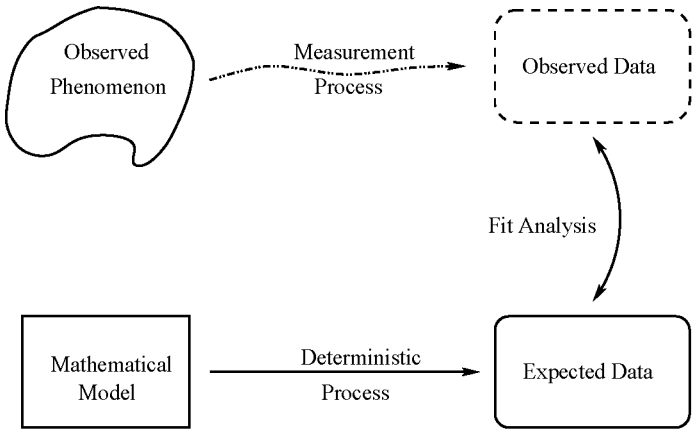

Figure 1.1: General Theory of Model Fit

the cornerstone of IRT, model fit was sought to be done since other important IRT concepts heavily depended upon its success.

Another important concept in IRT is the local independence (LID) (Ferrara, 1997; Chen & Thissen, 1997) of items, which stipulates that an item should not give information that could be used to answer any subsequent items (McCamey, 2002). Tests with strongly dependent items are to be avoided since their seemingly different questions are targeting the exact same area. Several indices have been developed to justify the LID assumption (Yen, 1994) and they rely on a good match between the model and the data. If the model fit is not assessed carefully and a LID index indicate significant departure from independence then there is no way to tell the reason. It can either be the misfit of the model or the presence of local dependence. Another

frequently utilized set of indices are developed to test against differential item functioning (DIF)(Dorans & Smith, 1993; Holland & Wainer, 1993). If one requires a reliable answer regarding DIF then the issue of model fit has to be taken seriously. These and other interesting indices are nicely reviewed by Ponoczny (2001).

A general model of fit theory is illustrated in Figure 1.1 which shows that there are two sources of data present: one represents the real world (or observed phenomenon) that one wishes to investigate and another which corresponds to a mathematical model that has been furnished and being thought to represent the real world. A good fit between these two objects is hoped for and needs justification. After obtaining representative data about the observed phenomenon the statistical analysis of fit can be carried out by applying one of the fit indices coupled with the mathematical analysis of the theoretical model. The selection of the right statistics is governed by the model of choice, its assumptions and the availability of indices.

The purpose of this dissertation is to introduce a new family of fit tests for the Rasch model. Figure 1.2 displays how the new technique (Monte Carlo method) applied in the new tests relates to the general notion of fit theory. The innovative aspect of the new tests is the creation of the test's distribution "on the fly" which makes these tests non-parametric, hence assumption free. The price one has to pay is that the exact mathematical calculation which is formidable even in the case of the Rasch model (let alone other, more complicated IRT models) is replaced by a Monte Carlo simulation method.

The structure of this dissertation is as follows. First the relevant terms used in the study will be introduced such as, IRT models, item characteristics function,

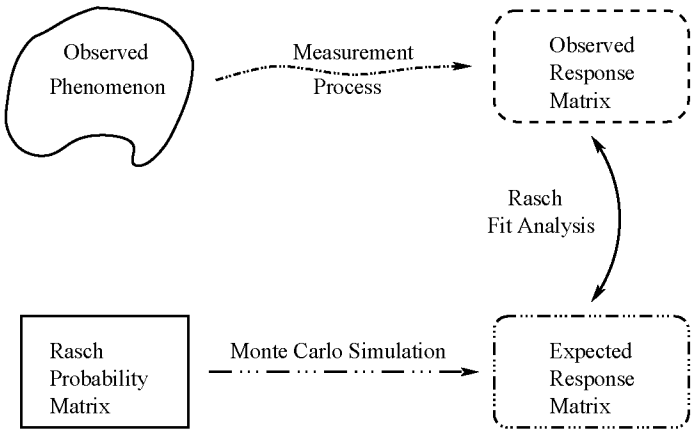

Figure 1.2: Model Fit in the Rasch Model

model assumptions, and the joint maximum likelihood estimation procedure. This will be followed by a detailed presentation of the fit problem in general. The heart of this introductory part is the discussion of the common method for approaching model fit, laying down the foundation of the newly proposed technique. This general point of view presented at this stage lights the road to the "Monte-Carlo solution" of the fit problem. Chapter 3 will give a thorough account of the geometrical picture underlying the typical residual-based fit analysis. This visual "aid" is mainly used to provide yet another motivation for the next section where the fundamental idea behind the new fit index is explained. Then comes the core of the dissertation, introducing the new family of fit indices in its entirety. This section also contains a theoretical comparison of the general, the residual-based, and the newly introduced fit tests, as well as a detailed description of the simulation method used. In Chapter

4 the performance of the new fit tests will be investigated under various conditions. Variables of interests are the size of response matrices, the distributional properties of test scores, the number of iterations necessary for obtaining stable p-values, and the association between the new test and the previously used residual-based fit indices in the Rasch model. These analyses are performed using raw and simulated data.

CHAPTER 2

IRT MODELS, ESTIMATION, AND MODEL FIT

2.1 Response Matrix

For the purpose of this study dichotomous models will be used exclusively, so the treatment of IRT extends only to these types of models.

Test data are usually summarized in the form of a two-dimensional data matrix, the *response matrix* $\mathcal{X}$, that contains responses of subjects to each test item. One row of the matrix contains the responses of a subject to all test items and a column reflects responses to an item from all subjects. As mentioned earlier, it is assumed that the responses are scored dichotomously. It means that the elements x_{ji} of $\mathcal{X}$ assume the value 1 or 0. Thus, x_{ji} is 1 if subject j responded correctly on item i, and 0 otherwise. The number of subjects is denoted by N and the number of items is by L.

The following 4×10 matrix is a possible data matrix for $N = 4$ subjects and $L = 10$ items.

$$\mathcal{X} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}. \tag{2.1}$$

Row marginal sums (r_j, $j = 1, \ldots, N$) represent the total score of each examinee (number of correct responses), and column marginal sums (s_i, $i = 1, \ldots, L$) represent the number of examinees responding correctly to a particular item:

$$r_j = \sum_{i=1}^{L} x_{ji}, \quad s_i = \sum_{j=1}^{N} x_{ji}. \tag{2.2}$$

Test theories, in their most pragmatic form, are developed to explain the response matrix $\mathcal{X}$. Their *raison d'etre* is the desire for a deeper understanding of the process that resulted in a very simple mathematical object $\mathcal{X}$. As argued later, the single challenge in this field is the huge number of possible outcomes. For a simple test situation with 8 items and 8 subjects the number of possible response matrices is 2^{64}, an incomprehensibly large number. It will be seen later, how this simple problem is able to control and restrict the treatment of fundamental questions in item response theory.

2.2 Classical Test Theory and Item Response Theory

For many decades test theory was dominated by a theory now referred to as *classical test theory* (CTT) (Hopkins, 1998, Allen & Yen, 1979). In the framework of this theory one assumes that subjects have a single trait (ability) that influences the response given to an item, which is assumed to possess certain difficulty. These parameters are estimated by the row and column marginal sums (or by the corresponding p-values), respectively. The model postulates, what kind of answers a pair of items can receive. That is, for an item pair (i_1,i_2) with $s_{i_2} < s_{i_1}$ the acceptable answer patterns are

$(1,1)$, $(0,0)$, $(1,0)$. In other words one is not allowed to give a correct answer to an item after missing an easier one.

Unfortunately, for CTT, in a real test-taking situation the existence of perfect Guttmann patterns (as they came to be called) are extremely rare. The appearance of an unexpected answer pattern $(0,1)$, in accordance with CTT, can only be attributed to measurement error. On the other hand, it is so frequent in practice, that one simply cannot achieve measurement without a huge measurement error. This drawback of CTT puzzled social scientists for a long time, but then a better alternative, called item response theory slowly emerged and has gradually gained popularity.

2.3 Response Probability Matrix

The fundamental idea of item response theory is that it allows any response pattern, as it only prescribes their *probability*. More precisely, item response theory postulates the existence of a *response probability* matrix $\mathcal{P}$ of the exact same size as $\mathcal{X}$. The elements of $\mathcal{P}$ are denoted by P_{ji}. The theory also states that the probability of a correct answer to item i by subject j is exactly P_{ji}.

Item response theory has a great flexibility compared to CTT. This statement can be made more transparent if we realize that item response theory actually contains CTT. One obtains CTT from IRT by choosing a very degenerate response probability matrix: $\mathcal{P} = \mathcal{X}$. This works only for the dichotomous model in this form. This choice means that a subject's answer is strictly determined by whether the subject's ability is larger than what is required for the correct answer or not. If the ability is larger, then

the probability of correct answer is 1, otherwise zero. For a test with many reversed Guttmann patterns the model becomes self contradictory. It predicts a correct answer even for too difficult items (for items that are more difficult than those that received incorrect answers). Also, in this case $\mathcal{P}$ is very rigid and extremely sensitive for even the smallest measurement error. Much more importantly, $\mathcal{P}$ is not determined by the ability of the subject and the difficulty of the item, so these notions simply do not gain real existence in CTT.

Item response theory becomes really powerful by the introduction of special *item characteristic functions*. IRT borrows the notion of subject ability and item difficulty from classical test theory and assumes that P_{ji} depends on them in a specific manner. More precisely, one assumes the existence of two collections of vector-valued random variables (Θ_j), $j = 1, \ldots, N$ and (Δ_i), $i = 1, \ldots, L$ the subject ability and item difficulty, respectively. Then one chooses a function $f = f(\Delta, \Theta)$ of these variables and prescribes that

$$P_{ji} = f(\Delta_i, \Theta_j), \tag{2.3}$$

for item i and subject j.

Now, the response probability matrix $\mathcal{P} = (P_i)$ is determined fully by ability and difficulty parameters, so using smart test-designing strategies it is possible to compare abilities of subjects who have never taken the same test. This is the underlying theme of one of the earliest books on the Rasch model by Wright and Stone (1979) which made a huge contribution to the work of psychometric firms performing large-scale assessment with tens of thousands of students throughout the United States.

To many, item response theory means a specific choice of f. In the next section the most important item response theory models will be reviewed, with respect to their item characteristic functions.

2.4 Logistic Models

In accordance with the model choice of this study only *unidimensional* models will be covered namely, models with a single random variable for each subject. The main distinction among models then is made by the ICF and the number of difficulty variables introduced. An intuitive step to reduce the collection of possible ICFs is taken by allowing only functions which are monotonic with respect to the ability variable. Models with this property are called *monotonic* item response theory models.

First, note that by taking *any* monotonic real function of one variable f with lower asymptote 0, upper asymptote 1 and with slope $f'(0) = 1$, one can create a family of 2, 3 or 4 parameter models by introducing the 4 parameter unidimensional, monotone model with $\Delta = (a, b, c, d)$ and $\Theta = (\vartheta)$

$$F(a, b, c, d; \vartheta) = c + (d - c)f(a(b - \vartheta)). \tag{2.4}$$

Taking $d = 1$ gives the 3 parameter model from which one obtains the 2 parameter model by requiring $c = 0$. If, in addition, $a = 1$ then one obtains a 1 parameter model.

A very popular family of IRT models is represented by the logistic models. The family borrows its name from the logistic function (Birnbaum, 1968; Baker, 1992, van der Linden & Hambleton, 1997)

$$f(\vartheta) = \frac{1}{1 + e^{-\vartheta}} \tag{2.5}$$

playing the role of the ICF in these models. The simplest model is the one-parameter logistic model, a.k.a. the Rasch model. The detailed account of this model is postponed to the next section.

The two parameter logistic model (2PLM) utilizes two item parameters. Its ICF is given by

$$f(a, b; \vartheta) = \frac{1}{1 + e^{a(b-\vartheta)}}. \tag{2.6}$$

Routine calculation shows that the slope of the tangent line to f (as a function of b) along the set $\{b = \vartheta\}$ is a. That is why one refers to a as the *slope*. Since both a and b are attributed to the item, these parameters together represent item difficulty.

One sometimes wants to have more freedom and introduces a third parameter to the model yielding the three parameter logistic model (3PLM). The ICF of 3PLM is

$$f(a, b, c; \vartheta) = c + \frac{1 - c}{1 + e^{a(b-\vartheta)}}. \tag{2.7}$$

The parameter c ($0 \leq c < 1$), being the lower asymptote of f, represents the probability of giving a correct answer with even extremely low ability. c is called the guessing factor, but sometimes this naming is rejected.

The 4PLM can be derived by introducing an upper asymptote for the ICF:

$$\hat{f}(a, b, c, d; \vartheta) = c + \frac{d - c}{1 + e^{a(b-\vartheta)}}. \tag{2.8}$$

d ($c < d \leq 1$) stands for the probability of achieving the correct answer with extremely high ability, usually called the ceiling effect. Parameter estimation procedures frequently run into trouble due to the large number of parameters to be estimated.

2.5 Rasch Model

The Rasch model utilizes 2 variables: item difficulty and person ability. The difficulty of item i and the ability of person j are denoted by δ_i and ϑ_j, respectively.

Under the Rasch model (Rasch, 1960, Wright & Stone, 1979; Wright & Masters, 1982) the conditional probability of subject j scoring on item i correctly (P_{ji}) or incorrectly (Q_{ji}) given the ability of the subject (ϑ_j) and the difficulty of the item (δ_i) is defined by the following functions:

$$P_{ji} := Prob(x_{ji} = 1 \mid \vartheta_j, \delta_i) = \frac{e^{\vartheta_j - \delta_i}}{1 + e^{\vartheta_j - \delta_i}} = \frac{1}{1 + e^{\delta_i - \vartheta_j}}, \tag{2.9}$$

$$Q_{ji} := Prob(x_{ji} = 0 \mid \vartheta_j, \delta_i) = 1 - P_{ji} = \frac{1}{1 + e^{\vartheta_j - \delta_i}}. \tag{2.10}$$

Function 2.9 is called the *Rasch item characteristic function* (or *curve*), Figure 2.1 shows the Rasch ICF.

The *Rasch* or *response probability matrix* $\mathcal{P}$ is the matrix with elements P_{ji} (Equation 2.9).

The conditional probability of one matrix element x_{ji} is then given by

$$P(x_{ji}) := \text{Prob}(x_{ji} \mid \delta_i, \vartheta_j) := \begin{cases} P_{ji} & \text{if} \quad x_{ji} = 1, \\ Q_{ji} = 1 - P_{ji} & \text{if} \quad x_{ji} = 0. \end{cases} \tag{2.11}$$

Also,

$$Q(x_{ji}) := 1 - P(x_{ji}). \tag{2.12}$$

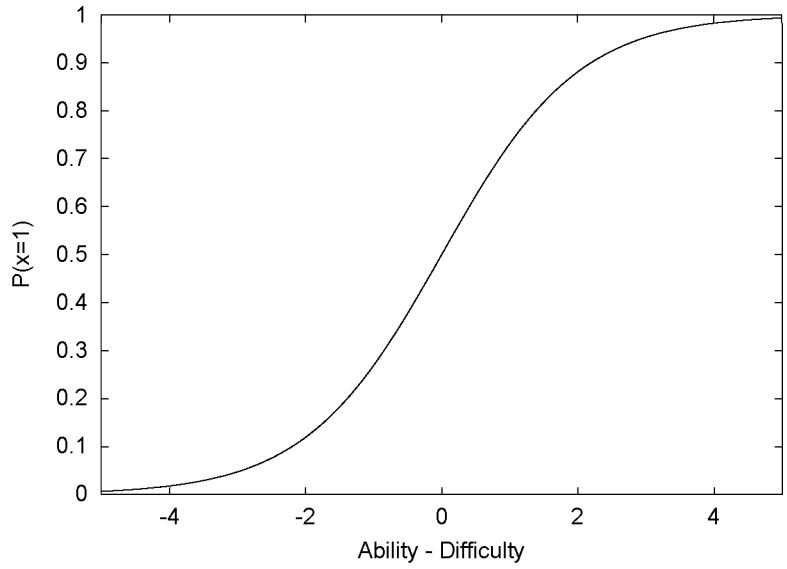

Figure 2.1: The Rasch Item Characteristic Function

It is worthwhile to note that the difficulty and ability values are real numbers without measurement unit. Their real meaning becomes apparent only when one introduces the conditional probability of each element of the item response matrix as in Equation 2.11. It is customary to say that the scale of ability and difficulty is a logarithmic measure called *logit* which expresses that they are on the same scale.

There are two important properties of the Rasch model which are already built in. One of them is the fact that the Rasch ICF is a monotone increasing function. It reflects the expectation that the probability of correctly answering to an item increases with ability. There is always a better chance for correct response with higher ability than with a lower one.

The second assumption that has already been utilized is that there is only one "trait" influencing the answer of a subject. This means more precisely, that there exists a family of random variables $\Theta = (\vartheta_j),\ j = 1, \ldots, N$ (the ability or trait). The ability of subject j is ϑ_j. One does not have to make any assumptions on the distribution of Θ, though. The presence of only one ability variable in the model implies that the Rasch model is *unidimensional*. Note that even though unidimensionality is formulated in terms of ability it is a feature of the test not the subjects taking it. In other words, it is a feature of the model used to explain the process of testing.

The presence of the other family of random variables the difficulty, $\Delta = (\delta_i),\ i = 1, \ldots, L$, also has to be considered. The difficulty of item i is δ_i.

What is crucial for the theory is the conditional probability of the entire item response matrix. For this, knowing the conditional probability P_{ji} is not sufficient. An additional condition is needed. The simplest choice is that the conditional probability of the response matrix is the product of the conditional probabilities of its elements. This definition requires the statistical independence of the matrix elements. It is expressed in the following equality

$$L(\mathcal{X}) := \text{Prob}(\mathcal{X} \mid (\vartheta_j, \delta_i)_{1 \le i \le L, 1 \le j \le N}) = \prod_{(x_{ji}) = \mathcal{X}} P(x_{ji}). \tag{2.13}$$

Now everything is ready to formulate the Rasch model. Recall that the goal of the Rasch model is to explain the response matrix $\mathcal{X}$.

Definition 1 *The Rasch Model for a response matrix $\mathcal{X}$ is given by the following:*

- *Two families of random variables Δ and Θ (difficulty and ability);*

- *The conditional probability $P(x_{ji})$ of the matrix element x_{ji}, given by (2.11);*
- *The conditional probability $L(\mathcal{X})$ of $\mathcal{X}$, given by (2.13).*

2.6 Comparison of Logistic IRT Models

In this section arguments defending the Rasch model over other IRT models will be introduced. Given its position as the simplest logistic model, the Rasch model is frequently judged pre-mature and un-suitable for good-fitting analysis. It is a fair criticism since the Rasch model is strictly contained in the 2PLM and as such, will never allow a better fit than the 2PLM, let alone the 3PLM or the 4PLM.

In the measurement society there has been a misunderstanding regarding the measurement necessity for conjoint additivity. According to Wright (1999) non-additive models like 2PLM and 3PLM models are mistaken as improvements over the 1PLM Rasch model. They introduce an item scaling parameter a to estimate discrimination and a lower asymptote parameter c to estimate a guessing level for each item. The negative consequences of applying these models are evident. For example, bias in person measures is significant (Stocking, 1989), and estimation procedures in 3 PLM can drift out of bounds (Swaminathan, 1983).

Wright (1999) also claims that these models destroy additivity by introducing item discrimination as a new parameter, while a and b cannot be estimated independently.

Another criticism of the many-parameter models is that their variables are not uniquely defined by ICFs that cross, for example, slopes differ due to differing discriminations, or asymptotes differ due to differing guessing parameters. This yields

crossing curves and the hierarchy of relative item difficulty can change at every ability level. In this situation there will be no criterion definition of the variable of interest.

"To construct measures we require orderly, cooperating, non-crossing curves like the Rasch curves. This means that we must take the trouble to collect and refine data so that they serve this clearly defined purpose, so that they approximate a stochastic Guttmann scale" (Wright, 1999, p.97).

Yet another motivation behind fit analysis is the need for clarifying the correspondence of the different IRT models in terms of model fit. One sees immediately that the fit of the Rasch model can never be better than the fit of the other members of the logistic family. However, knowing the fit of the Rasch model with a certain degree of reliability greatly helps researchers in deciding between the appealingly simple Rasch model and the conceptually not so well founded and mathematically questionable 2, 3, or 4 PLMs.

2.7 Joint Maximum Likelihood Estimation

Estimation procedures make parameter estimates available through an iteration procedure. The joint maximum likelihood estimation (JMLE) (Baker, 1992; Fisher, 1981; Myung, in press) procedure considers difficulties and abilities on the same footing. By definition, JMLE finds parameter estimates such that the likelihood function $L(\Delta, \Theta)$ corresponding to the response matrix $\mathcal{X}$ is maximal.

In practice it is much more convenient to use the log-likelihood function

$$\mathcal{L} := \log(L(\mathcal{X})) = \sum_{(x_{ji})=X} \log\left(\mathrm{P}(x_{ji})\right) = -\sum_{(x_{ji})=\mathcal{X}} \log\left(1 + e^{(2x_{ji}-1)(\delta_i - \vartheta_j)}\right). \tag{2.14}$$

Note, that finding the maximum place of $L(\mathcal{X})$ and that of $\mathcal{L}$ are equivalent since the logarithm is a strictly monotone increasing function.

Locating the maximum of $\mathcal{L}$ is usually done by finding the zeros of the derivative $D\mathcal{L}$ of $\mathcal{L}$, since, if $\mathcal{L}$ has a unique maximum then it occurs at the zero of $D\mathcal{L}$. This maximum problem is unsolvable by analytic methods (especially for large tests), therefore a numerical method is called for. Frequently, the Newton-Raphson algorithm is used to find the zeros of $D\mathcal{L}$.

Let us recall the Newton-Raphson algorithm for finding the zeros of the first derivative matrix $D\mathcal{L}$ (Kress, 1998, p. 102). First, let us choose an arbitrary initial point

$$x_0 = (\delta_1^0, \delta_2^0, \ldots, \delta_L^0, \vartheta_1^0, \vartheta_2^0, \ldots, \vartheta_N^0) \tag{2.15}$$

for the iteration. The iteration scheme is then given by

$$x_{n+1} = x_n - \left(D^2\mathcal{L}(x_n)\right)^{-1} \cdot D\mathcal{L}(x_n), \tag{2.16}$$

where

$$x_n = (\delta_1^n, \delta_2^n, \ldots, \delta_L^n, \vartheta_1^n, \vartheta_2^n, \ldots, \vartheta_N^n) \tag{2.17}$$

denotes the nth approximation of zero, and $(D^2\mathcal{L}(x_n))^{-1}$ is the inverse of the second derivative matrix of $\mathcal{L}$ evaluated at x_n.

Due to the special form of the second derivative matrix this procedure has a unique solution. In other words, the iteration scheme converges to a unique solution no matter what the initial value of x_0 is. This behavior seems to be a special feature of the Rasch model that is not, in general, shared by other IRT models.

Let us now embark on the calculation of the derivatives of $\mathcal{L}$ involved in the computation. The components of the first derivative function are given by

$$D\mathcal{L}_i = \frac{\partial \mathcal{L}}{\partial \delta_i} = -\sum_{j=1}^{N} \frac{\pm e^{\pm(\delta_i - \vartheta_j)}}{1 + e^{\pm(\delta_i - \vartheta_j)}} = \sum_{j=1}^{N} \frac{\mp 1}{1 + e^{\mp(\delta_i - \vartheta_j)}}, \quad 1 \le i \le L, \tag{2.18}$$

$$D\mathcal{L}_{L+j} = \frac{\partial \mathcal{L}}{\partial \vartheta_j} = -\sum_{i=1}^{L} \frac{\mp e^{\pm(\delta_i - \vartheta_j)}}{1 + e^{\pm(\delta_i - \vartheta_j)}} = \sum_{i=1}^{L} \frac{\pm 1}{1 + e^{\mp(\delta_i - \vartheta_j)}}, \quad 1 \le j \le N, \tag{2.19}$$

where $\pm$ (resp. $\mp$) stands for $2x_{ji} - 1$ (resp. $1 - 2x_{ji}$).

The non-zero elements of the second derivative of $\mathcal{L}$ are

$$D^2\mathcal{L}_{ii} = \frac{\partial^2 \mathcal{L}}{\partial \delta_i^2} = \sum_{j=1}^{N} \frac{-e^{\mp(\delta_i - \vartheta_j)}}{\left(1 + e^{\mp(\delta_i - \vartheta_j)}\right)^2} = -\sum_{j=1}^{N} P_{ji} Q_{ji}, \quad 1 \le i \le L, \tag{2.20}$$

$$D^2\mathcal{L}_{L+j,L+j} = \frac{\partial^2 \mathcal{L}}{\partial \vartheta_j^2} = \sum_{i=1}^{L} \frac{e^{\mp(\delta_i - \vartheta_j)}}{\left(1 + e^{\mp(\delta_i - \vartheta_j)}\right)^2} = -\sum_{i=1}^{L} P_{ji} Q_{ji}, \quad 1 \le j \le N, \tag{2.21}$$

$$D^2\mathcal{L}_{i,L+j} = \frac{\partial^2 \mathcal{L}}{\partial \delta_i \partial \vartheta_j} = \frac{-e^{\mp(\delta_i - \vartheta_j)}}{\left(1 + e^{\mp(\delta_i - \vartheta_j)}\right)^2} = P_{ji} Q_{ji}, \tag{2.22}$$

$$D^2\mathcal{L}_{L+j,i} = \frac{\partial^2 \mathcal{L}}{\partial \vartheta_j \partial \delta_i} = D^2\mathcal{L}_{i,L+j} = P_{ji} Q_{ji}, \tag{2.23}$$

where in the last two equations $1 \leq i \leq L$ and $1 \leq j \leq N$. Note that the equality $P(x_{ji})Q(x_{ji}) = P_{ji}Q_{ji}$ is used .

It is transparent that

$$\sum_{j=1}^{N} D^2\mathcal{L}_{L+j,i} = -D^2\mathcal{L}_{ii} \quad \text{and that} \quad \sum_{i=1}^{L} D^2\mathcal{L}_{L+j,i} = D^2\mathcal{L}_{L+j,L+j}. \tag{2.24}$$

It is customary to quote some assumptions (independence of the ability and difficulty estimates) of the Rasch model which would imply that $D^2\mathcal{L}_{L+j,i} = 0$ for all i and j. It is clear from the equalities in Equation 2.24 that this would, in turn, imply that $D^2\mathcal{L}_{ii} = D^2\mathcal{L}_{L+j,L+j} = 0$, as well. This is not only a contradiction but would also make the entire calculation meaningless since it would imply that $D^2\mathcal{L} = 0$ *everywhere.*

Note, however, that in the generic case the off-diagonal matrix elements are much smaller than the diagonal ones. This is true for test situations when the number of subjects N and the number of items L are large. That is why off-diagonal elements are omitted from the calculation noting that for relatively small tests or for non-generic situations this is a very questionable approximation.

After the simplification, $D^2\mathcal{L}$ becomes an everywhere strictly negative definite matrix, implying that the zero of its first derivative is really a maximum value of the likelihood function.

Recognize that, due to the special form of the Rasch ICF it is invariant to translations. That is,

$$f_{Rasch}(\delta + s, \vartheta + s) = f_{Rasch}(\delta, \vartheta) \tag{2.25}$$

for all real numbers s. To resolve this ambiguity in the model an additional condition has to be imposed. This is usually done by requiring that the mean of the item difficulties is 0. This is achieved by performing the following transformation to the estimates ($\overline{\delta} = \frac{1}{L}\sum_{i=1}^{L}\delta_i$ is the mean of $(\delta_i)_{i=1}^{L}$):

$$\begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_L \\ \vartheta_1 \\ \vartheta_2 \\ \vdots \\ \vartheta_N \end{pmatrix} \mapsto \begin{pmatrix} \delta_1 - \overline{\delta} \\ \delta_2 - \overline{\delta} \\ \vdots \\ \delta_L - \overline{\delta} \\ \vartheta_1 \\ \vartheta_2 \\ \vdots \\ \vartheta_N \end{pmatrix} \tag{2.26}$$

In a Newton-Raphson scheme this extra condition needs to be taken into account by adjusting the mean item difficulty estimates to zero after each iteration step. Denoting by $*$ the un-adjusted estimates, the details of the JMLE procedure are as

follows:

$$x_0^* = \begin{pmatrix} \delta_1^{0*} \\ \delta_2^{0*} \\ \vdots \\ \delta_L^{0*} \\ \vartheta_1^0 \\ \vartheta_2^0 \\ \vdots \\ \vartheta_N^0 \end{pmatrix} \rightarrow x_0 = \begin{pmatrix} \delta_1^0 \\ \delta_2^0 \\ \vdots \\ \delta_L^0 \\ \vartheta_1^0 \\ \vartheta_2^0 \\ \vdots \\ \vartheta_N^0 \end{pmatrix} \Rightarrow x_1^* = \begin{pmatrix} \delta_1^{1*} \\ \delta_2^{1*} \\ \vdots \\ \delta_L^{1*} \\ \vartheta_1^1 \\ \vartheta_2^1 \\ \vdots \\ \vartheta_N^1 \end{pmatrix} \rightarrow x_1 = \begin{pmatrix} \delta_1^1 \\ \delta_2^1 \\ \vdots \\ \delta_L^1 \\ \vartheta_1^1 \\ \vartheta_2^1 \\ \vdots \\ \vartheta_N^1 \end{pmatrix} \Rightarrow \ldots \quad (2.27)$$

where $\rightarrow$ symbolizes the mean adjustment (Equation 2.26) and $\Rightarrow$ stands for the Newton-Raphson iteration step (Equation 2.16).

The iteration stops when the norm[1] of the difference

$$\|x_{n-1} - x_n\| \quad (2.28)$$

between two consecutive parameter estimate vectors is smaller than a prescribed number. This number is usually called the margin of iteration. In practice this number is chosen to be around 0.001 and in any computer program this can be set by the researcher. The last element x_n in the iteration scheme is the estimate of the parameters sought by the researcher.

[1] The usual Euclidean norm is defined as

$$\|x\| := \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Note that from a practical point of view the estimates obtained by this method can be considered exact solutions to the mathematical problem of finding the zeros of $D\mathcal{L}$. By adjusting the iteration margin appropriately we can achieve any prescribed precision. If estimation procedures that are based on the same algorithm give different estimates then that cannot be attributed to the fact that they are estimates. These estimates remain estimates in the sense that the measurement procedure itself carries an inevitable measurement error.

2.7.1 Data "Cleaning"

There is a very important computational, as well as conceptual issue that has not yet been discussed. It concerns the fact that the estimation procedure cannot converge if there are rows or columns with all 0s or all 1s (see Wright & Masters, 1979). Let us assume that the first row consists of only 1s, that is the first subject gave correct answer to all items. Let us choose *any* set of estimates $(\delta_1, \ldots, \delta_L, \vartheta_1, \ldots, \vartheta_N)$. The contribution $\mathcal{L}_1$ to the log-likelihood $\mathcal{L}$ coming from the first row is

$$\mathcal{L}_1 := -\sum_{i=1}^{L} \log(1 + e^{\delta_i - \vartheta_1}). \tag{2.29}$$

Adding any positive number to ϑ_1 will increase $\mathcal{L}_1$ and in turn $\mathcal{L}$. This means that the maximum of the log-likelihood can only be achieved at a point where the ability estimate of the first respondent is ∞. The same reasoning yields that in case of a constant row or column the estimate vector will lie outside any bounded region.

The conceptual explanation is that those subjects that answer all questions are too smart for the test. These subjects cannot contribute to the measurement targeted by

the specific test and they have to be discarded from the analysis. By the same token, those subjects that answer all questions wrong do not attribute to the measurement process either. Therefore, data "cleaning" is not an attempt to artificially improve fit. The removal of rows and columns with constant values has solid computational and conceptual reasons.

This procedure can be easily demonstrated by looking at a moderately sized response matrix but requires extensive computing time for larger matrices. This "cleaning" procedure is comparable in running time to the entire estimation procedure.

The following example will present the steps of a general "data cleaning" procedure. The response matrix to start with is the one introduced in Chapter 2 (2.1) with size of 4×10. After removing two columns, two rows and again one row the final "cleaned" response matrix becomes a 3×6 matrix.

The steps are as follows:

$$\mathcal{X} = \begin{pmatrix} \mathbf{1} & 1 & 1 & 0 & 1 & 1 & 0 & \mathbf{0} & 0 & 0 \\ \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} & 0 & 0 \\ \mathbf{1} & 1 & 1 & 1 & 0 & 1 & 1 & \mathbf{0} & 1 & 0 \\ \mathbf{1} & 1 & 0 & 0 & 0 & 1 & 1 & \mathbf{0} & 0 & 1 \end{pmatrix} \rightarrow \tag{2.30}$$

$$\rightarrow \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} \mathbf{1} & 1 & 0 & 1 & \mathbf{1} & 0 & 0 & 0 \\ \mathbf{1} & 1 & 1 & 0 & \mathbf{1} & 1 & 1 & 0 \\ \mathbf{1} & 0 & 0 & 0 & \mathbf{1} & 1 & 0 & 1 \end{pmatrix} \rightarrow$$

$$\rightarrow \mathcal{X}_{\text{clean}} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

2.8 Rasch Model Fit Analysis

The importance of model fit can not be overstated. No matter how consistent and well established a theory is if the data are not explained an alternative has to be found. Data fit simply means the extent to which the underlying social and/or psychological process, represented by the data, is explained by the model. If the model is able to reproduce the observed data with high probability then one can say that the data fit the model (see Figure 1.1). The higher the probability of the observed data the more one believes in the model. Since perfect fit can rarely be achieved one has to develop an array of tools to measure the "distance" of the observed data from the data created by the model. The quantifications of this distance are called *fit indices.*

Models that can stand several decades of investigation and almost always display good fit constitute the foundations of educational measurement. Without well-fitting models educational and psychological measurement could not exist.

Many advocates of the Rasch model argue, that the Rasch model is not only simple and beautiful, but due to the conceptual difficulties of other item response theory models, it is essentially the only one that could be used in a sensible manner. The biggest criticism is that the simplicity of the model also results in the worst fit (at least in the logistic family of IRT models). For, the exact form of the other logistic models (2PLM, 3PLM, 4PLM) ICF reveals that the maximum likelihood estimation procedure for the Rasch model is the same as the maximum likelihood estimation procedure for the higher logistic models restricted to the subset of the variable space

where the additional parameters are set to appropriate constants. Having more freedom in the choice of the parameter estimates clearly results in a better (or at least not worse) fit.

Arguably, to see how "badly" the Rasch model fits the fit analysis must be taken very seriously. When the fit analysis introduces extra assumptions and fails, then not only the model, but the extra assumptions become suspicious.

By no surprise, the theory of fit in IRT is not much younger than the theory itself. Wright & Stone (1979) and Wright & Masters (1982) introduced the mean-square (MNSQ) fit indices. Practically, these are distances between the response probability matrix $\mathcal{P}$ and the response matrix $\mathcal{X}$ with respect to specific weights.

The w-distance of two matrices A and B of the same size is given by the square root of the function

$$D_w^2(A, B) := \sum_{ij} (A_{ij} - B_{ij})^2 w_{ij}, \tag{2.31}$$

where w is the *weight* matrix satisfying: $w_{ij} > 0$ for all i, j.

A residual fit index is usually of the form $D_w^2(\mathcal{X}, \mathcal{P})$ with some special weight w.

2.8.1 Mean-Square and Log-likelihood Tests

The first item-fit statistics for the Rasch model proposed by Wright and Panchapakesan (1969) were based on person raw score groups which focused on the difference between the observed and expected score for a group of persons with the same raw score on a test (Smith, 1995). Subsequent developments were based on the item and person residuals and were expressed in a chi-square form. Then, this chi-square

statistic were converted to a mean-square by dividing by the number of degrees of freedom.

The total MNSQ outfit is defined as

$$\text{MNSQ}_{out} = \frac{1}{NL} \sum_{i=1}^{L} \sum_{j=1}^{N} \frac{(x_{ji} - P_{ji})^2}{P_{ji} Q_{ji}}, \tag{2.32}$$

MNSQ infit is defined as

$$\text{MNSQ}_{in} = \frac{\sum_{i=1}^{L} \sum_{j=1}^{N} (x_{ji} - P_{ji})^2}{\sum_{i=1}^{L} \sum_{j=1}^{N} P_{ji} Q_{ji}}. \tag{2.33}$$

According to Wright and Stone (1979) and Wright and Masters (1982), MNSQ outfit follows an approximate χ^2 distribution (after multiplying by NL) with NL degrees of freedom.

These statistics originally were evaluated as fit mean-squares in BICAL, an early Rasch calibration program, with expected value of 1. Unfortunately, the critical values for detecting misfit depend on the the degrees of freedom so they will vary from sample to sample (Smith, 1995). Due to the difficulty involved in handling the χ^2 distribution the MNSQ score were transformed to an approximate unit normal distribution using the Wilson-Hilferty transformation (Patel & Read, 1996).

More precisely one defines the standard z score belonging to MNSQ outfit as follows:

$$\text{MNSQ}_{out,zstd} := \sqrt{\frac{9NL}{2}} \left(\sqrt[3]{MNSQ} - 1 + \frac{2}{9NL} \right). \tag{2.34}$$

For infit one finds the standardized value by

$$\mathrm{MNSQ}_{in,zstd} := \frac{3\left(\sqrt[3]{\mathrm{MNSQ}_{in}} - 1\right)}{q} - \frac{q}{3}, \tag{2.35}$$

where q is the standard deviation of infit given by

$$q = \frac{\sum_{i=1}^{L}\sum_{j=1}^{N} P_{ji}Q_{ji}(P_{ji} - Q_{ji})^2}{\sum_{i=1}^{L}\sum_{j=1}^{N} P_{ji}Q_{ji}} \tag{2.36}$$

(see Wright & Masters. 1982 p. 100).

Based on the assumptions, both MNSQ_{zstd} follow a standard normal distribution $N(0,1)$.

Levine & Rubin (1979) applied the log-likelihood function to assess fit.

$$\mathcal{L}_o := \mathcal{L}(\mathcal{X}) := \log(L(\mathcal{X})) = \sum_{i=1}^{L}\sum_{j=1}^{N} x_{ji}\log P_{ji} + (1 - x_{ji})\log(1 - P_{ji}). \tag{2.37}$$

Drasgow, Levine, & Williams (1985) have used the standardized version $\mathcal{L}_z$ of this index, that is asymptotically standard normally distributed.

$$\mathcal{L}_z = \frac{\mathcal{L}_o - \mathcal{E}(\mathcal{L}_o)}{\sigma(\mathcal{L}_o)}, \tag{2.38}$$

where

$$\mathcal{E}(\mathcal{L}_o) := \sum_{i=1}^{L}\sum_{j=1}^{N} P_{ji}\log P_{ji} + (1 - P_{ji})\log(1 - P_{ji}) \tag{2.39}$$

and

$$\sigma^2(\mathcal{L}_o) := \sum_{i=1}^{L}\sum_{j=1}^{N} P_{ji}(1 - P_{ji})\log^2\left(\frac{P_{ji}}{1 - P_{ji}}\right). \tag{2.40}$$

There is an ongoing battle between these fit indices (Li & Olejnik, 1997; Linacre, 1997) even though both of them are perfect candidates for assessing model or person/item fit. The main drawback of both is that they use extra assumptions about their distributions that are not always satisfied.

2.8.2 General Theory of Model Fit

In this section the general theory of item response theory model fit is discussed. The generality of the presentation will allow pinpointing where the particular fit indices enter the picture and how their deficiency can be remedied.

In the framework of dichotomous IRT, for even a moderately large response matrix, the number of possible response matrices is already incomprehensibly large. The number of dichotomous response matrices of size $N \times L$ is 2^{NL}. In this humongous set one has to calculate the probability of each element (under the assumption that the model holds). As will be shown, this is practically impossible, making the theory of model fit a subtle problem.

To be more specific, let there be given a response matrix $\mathcal{X}$ of size $N \times L$. It is assumed from now on that the estimation procedure has been carried out and the parameter estimates have been obtained. These estimates give rise to the response probability matrix $\mathcal{P}$ via Equation 2.9. Then, one considers the collection of all response matrices $M_{resp}(N, L)$ with size $N \times L$. The set $M_{resp}(N, L)$ is defined to be the collection of N by L matrices with entries 0 and 1.

Let $\mathcal{Y}$ be an element in $M_{resp}(N, L)$. The probability required for the hypothesis testing concerning model fit is the conditional probability

$$\text{Prob}(\mathcal{Y} \mid \mathcal{P}) := \prod_{ji} (y_{ji} P_{ji} + (1 - y_{ji}) Q_{ji}). \tag{2.41}$$

(Note the subtle difference between $L(\mathcal{X})$ and $\text{Prob}(\mathcal{Y} \mid \mathcal{P})$.)
This probability measure $\mathfrak{p}(\mathcal{P})$ on $M_{resp}(N, L)$

$$\mathfrak{p} := \mathfrak{p}(\mathcal{P}) : M_{resp}(N, L) \to [0, 1] : \ \mathcal{Y} \mapsto \text{Prob}(\mathcal{Y} \mid \mathcal{P}) \tag{2.42}$$

is the null distribution for the hypotheses testing.

Due to the enormity of the set $M_{resp}(N, L)$ for even moderate values of N and L it is absolutely hopeless to handle this null distribution in its entirety.

The next step in the hypotheses testing is to find the "tail" probability (or p-value) p_o of $\mathcal{X}$. This tail probability p_o can be defined as the sum of all conditional probabilities $\text{Prob}(\mathcal{Y} \mid \mathcal{P})$ satisfying $\text{Prob}(\mathcal{Y} \mid \mathcal{P}) < \text{Prob}(\mathcal{X} \mid \mathcal{P}) = L(\mathcal{X})$ over $\mathcal{Y} \in M_{resp}(N, L)$, that is

$$p_o := \sum_{\mathcal{Y}:\ \text{Prob}(\mathcal{Y} \mid \mathcal{P}) < \text{Prob}(\mathcal{X} \mid \mathcal{P})} \text{Prob}(\mathcal{Y} \mid \mathcal{P}). \tag{2.43}$$

By introducing

$$\mathcal{B}_L := \{\mathcal{Y} \ : \ \text{Prob}(\mathcal{Y} \mid \mathcal{P}) < \text{Prob}(\mathcal{X} \mid \mathcal{P})\} \tag{2.44}$$

one can rewrite p_o as an integral of the constant 1 function over $\mathcal{B}_L$ with respect to the probability measure $\mathfrak{p}$:

$$p_o := \mathfrak{p}(\mathcal{B}_L) := \int_{\mathcal{B}_L} 1 \ \mathrm{d}\mathfrak{p} \tag{2.45}$$

This form allows the formulation of the fit problem for a general fit index. For instance, the null probability of the MNSQ fit test is defined by using

$$\mathcal{B}_{\text{MNSQ}} := \{\mathcal{Y} \ : \ D^2(\mathcal{Y}, \mathcal{P}) \leq \text{MNSQ}(\mathcal{X}) \leq 1\} \cup \{\mathcal{Y} \ : \ D^2(\mathcal{Y}, \mathcal{P}) \geq \text{MNSQ}(\mathcal{X}) \geq 1\} \tag{2.46}$$

in place of $\mathcal{B}_L$ and defining the p-value by

$$p_o^{\text{MNSQ}} := \mathfrak{p}(\mathcal{B}_{\text{MNSQ}}) := \int_{\mathcal{B}_{\text{MNSQ}}} 1 \, \mathrm{d}\mathfrak{p}. \tag{2.47}$$

In other words, the null probabilities are always the measures of some specific sets defined by the particular fit index of choice. The difficulty in handling these indices is always the same, however. It lies in the practical impossibility of handling the probability distribution p by exact methods.

Note that the probability p_o in Equation 2.45 is the p-value of the L-test, which is the fit test using the likelihood function itself as the fit index. From the point of view of the maximum likelihood estimation this test seems to be the most natural choice.

The MNSQ and $\mathcal{L}_z$ tests with their probability normality assumptions show how practitioners make assumptions about their respective null distributions allowing a first approximation of fit analysis.

Having the null-probability p_o^V for some fit index V, the final task is to decide if it is larger or smaller than a prescribed threshold probability α (traditionally α lies in the range $[0.001, 0.1]$). When $p_o < \alpha$ then the null hypothesis is rejected and it is stated that there is not enough evidence to support that the data fit the model (or the model explains the data).

CHAPTER 3

A FAMILY OF NON-PARAMETRIC FIT INDICES

3.1 Statement of Problem

Unfortunately, both the MNSQ-based outfit and infit indices as well as the likelihood-based $\mathcal{L}_o$ and $\mathcal{L}_z$ use a strong assumption regarding their distributions. Namely, it is assumed that MNSQ statistics have a χ^2 distribution. It is customary then to convert it, using the Wilson-Hilferty transform, to MNSQ_{zstd} which is then assumed to have a standard normal distribution. This assumption is the *normality assumption of* MNSQ. $\mathcal{L}_z$ is also assumed to be standard normally distributed. In cases when these assumptions fail to hold, the appropriateness of MNSQ and of $\mathcal{L}_z$ are in danger. There are several studies targeting this issue (Li & Olejnik, 1997; Noonan, Boss & Gessaroli, 1992; Wright & Linacre, 1985; Smith, 1982). They all show evidence that the distribution of MNSQ statistics depart from normality.

Li and Olejnik (1997) conducted a simulation study for $\text{MNSQ}_{out,zstd}$ person fit with several test and misfit scenarios. For each scenario 50 replications were made resulting in a fit index distribution. Investigating the normality of these distributions using skewness and kurtosis they found that it significantly deviates from the normal distribution. A similar study was conducted by Noonan, Boss and Gessaroli (1992)

also yielding that $\text{MNSQ}_{out,zstd}$ person fit shows a great departure from normality with large skewness and extremely large variable kurtosis.

Similar results were found earlier by Smith (1982) and Wright and Linacre (1985). Wright and Linacre suggested raising the cut-off value for rejecting person fit to 3.0 (from the standard 2) or beyond to compensate for non-normality.

3.2 Rationale for the New Tests

A closer look at the MNSQ fit reveals that it is nothing else but a distance squared between the Rasch probability matrix $\mathcal{P}$ and the response matrix $\mathcal{X}$ with respect to specific weights. First, an overview of the elementary linear algebra involved is given.

In general, a vector $w = (w_i)_{i=1}^n \in \mathbb{R}^n$ is called a *weight* if its elements are all positive, that is $w_i > 0$ for all i. The distance between two vectors $x, y \in \mathbb{R}$ with respect to the weight w (or w-distance) is given by

$$D_w(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2 w_i}. \tag{3.1}$$

The name is justified, because D_w satisfies the conditions required in the definition of a distance (see e.g. Lang, 1986). That is, one has for all $x, y, z \in \mathbb{R}^n$

(i) $D_w(x,x) = 0 \Rightarrow x = 0$,

(ii) $D_w(x,y) = D_w(y,x)$,

(iii) $D_w(x,z) \leq D_w(x,y) + D_w(y,z)$.

From Equation 2.32 it is clear that MNSQ_{out} is the distance squared

$$\text{MNSQ}_{out} = D^2_{w_o}(\mathcal{P}, \mathcal{X}) \tag{3.2}$$

with weight

$$w_{o,ji} = \frac{1}{NL \cdot P_{ji}Q_{ji}}. \tag{3.3}$$

Similarly for the MNSQ infit (Equation 2.33) one has

$$\text{MNSQ}_{in} = D^2_{w_i}(\mathcal{P}, \mathcal{X}), \tag{3.4}$$

where the weight is

$$w_{i,ji} = \frac{1}{\sum_{i=1}^{L}\sum_{j=1}^{N} P_{ji}Q_{ji}}. \tag{3.5}$$

Note, that $w_{i,ji}$ is a constant vector, that is for the infit index one uses a constant weight.

Before starting the precise definition of the new fit indices let us give a short rationale for them.

Due to the reasons presented in the previous section a completely new direction has to be found for fit analysis to overcome those limitations. The use of predefined generic distributions has been an issue for a long time, not only in fit analysis but also in many other fields of statistical modeling.

The innovative idea for the new testing method is the creation of a sampling distribution representing the Rasch model and to create a second distribution that shows the characteristics of observed data. Both are created through computer simulation.

Furthermore, the original fit statistic $MNSQ_{outfit}$ and $MNSQ_{infit}$ will be utilized to some degree.

In generating the distributions the main idea is to use the distance interpretation of MNSQ. From this point of view, the goal of the fit test is to find the distance between $\mathcal{P}$ and $\mathcal{X}$ and then to assess how far we have gotten with the estimation procedure from the initial data $\mathcal{X}$. One has to be careful with the distance interpretation, though, as the expected value of this distance is not zero but 1.

The idea is to use computer simulated data to find out how much deviation of this distance from 1 should be considered too large. In the process the computer generates a new set of response matrices $\mathcal{Y}^k$, $k = 1, \ldots, K$ using the item difficulty and person ability estimates that actually specify the probability matrix P. (K is the number of simulated matrices)[1].

By construction, for these simulated matrices the Rasch model holds (note that the Rasch model has been used to generate them!). In other words, this simulated set shows the best case scenario for the test output from the Rasch point of view. Every one of them could have been a completely legitimate response matrix in a process governed by the Rasch model with ability and difficulty vectors given by the estimates obtained from the original response matrix. The task now is to find where $\mathcal{X}$, the original response matrix, is situated among these simulated matrices. *It is very important to note that by this simulation method a sample of response matrices is created in a way to make finding good fit for $\mathcal{X}$ as difficult as possible.*

[1] As can be seen later, it actually generates two sets ($\mathcal{Y}_1$ and $\mathcal{Y}_2$), because at some point of the hypotheses testing the independence of the simulated data is needed.

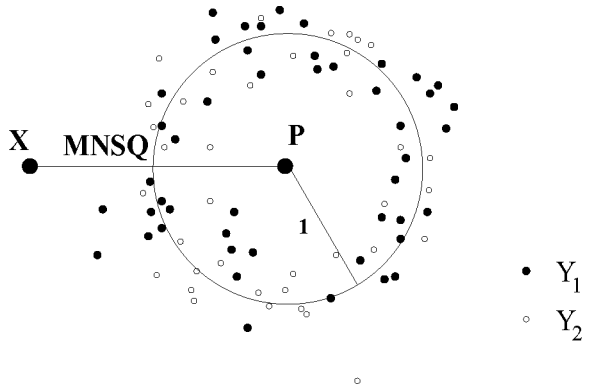

Figure 3.1: A non-fitting Rasch model

The new simulation technique proposed in this work is embodied in four different fit tests which will be discussed in the following sections in more detail. The tests aim to allocate $\mathcal{X}$ in relation to $\mathcal{P}$ which in turn has been replaced by $\mathcal{Y}^k$. The model/sampling distribution can be generated in two different ways with the aid of $\mathcal{Y}^k$. Also, the observed data will be represented by either the original MNSQ fit statistics or a distribution.

To give the reader a general idea the test that uses a sampling distribution and a fit statistic is illustrated in Figures 3.1 and 3.2 (a non-fitting and a fitting example, respectively).

If $\mathcal{X}$ sits outside the "cloud" of $\mathcal{Y}$s, then it is highly unlikely that $\mathcal{X}$ is the result of a process that can be described by the Rasch model (Figure 3.1). On the other hand, if $\mathcal{X}$ can be found inside the ring of simulated matrices then we tend to accept

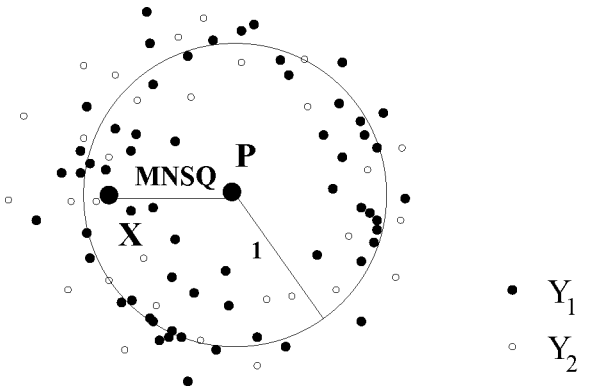

Figure 3.2: A fitting Rasch model

that the extremely complicated interaction between the test taker and the test can be actually understood with the aid of the Rasch model (Figure 3.2).

Figure 3.3 shows an empirical distribution of $D^2_{out}(\mathcal{Y}, \mathcal{P})$. For this run the number of simulated matrices chosen is $K = 1000$ and the number of bars is 50. This graph symbolizes a sampling distribution that has been simulated via a Monte Carlo method. It can be visualized as a projection of all simulated $\mathcal{Y}$ matrices on a 2 dimensional surface.

By now, it should be clear that this distribution will serve as a basis for the fit test. The p-value is obtained by finding the tail probability which is the area in the tail beyond the cut score. If this area is larger than the pre-defined α then fit is accepted; if not, the null hypothesis of model fit is rejected. In fit analyses, model fit is desired, so analysts are always looking for high p-values. The selection of α is dependent upon the purpose of the study, though.

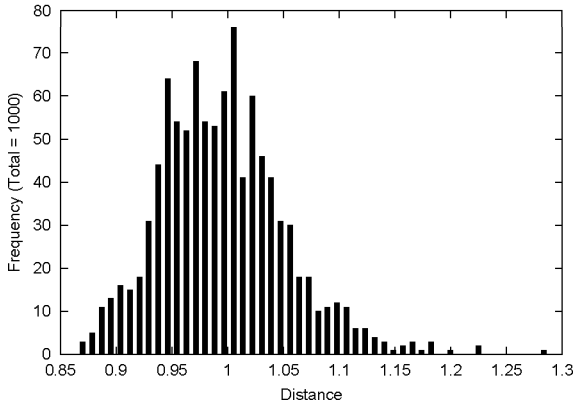

Figure 3.3: A $D^2_{out}(\mathcal{Y}, \mathcal{P})$ distribution

3.3 Simulation of Rasch Response Matrices

In this section the algorithm to simulate Rasch response matrices used in the new tests is described. The initial data are the Rasch probability matrix calculated from the item difficulty and person ability vectors, which are obtained from a suitable estimation procedure. The generation of a response matrix $\mathcal{Y}$ is done by generating each of its elements separately. To create a single response of subject j to item i the computer generates a random number r uniformly distributed on the interval $[0, 1]$. The response y_{ji} is then defined by

$$y_{ji} = \begin{cases} 1 \text{ if } r \leq P_{ji}, \\ 0 \text{ otherwise.} \end{cases} \tag{3.6}$$

where P_{ji} is the Rasch probability as in Equation 2.9. The parameter estimates are used through P_{ji}, of course.

It is clear from the construction, that for the expected value of $\mathcal{Y}$ one has the following formula

$$\mathcal{E}(\mathcal{Y}) = \mathcal{P} \quad \text{or} \quad \mathcal{E}(y_{ji}) = P_{ji} \; \forall (i,j). \tag{3.7}$$

The heart of the Monte Carlo fit test is the very simple consequence of this simulation method:

Observation: *The relative frequency of a response matrix $\mathcal{Y} \in M_{resp}(N, L)$ in the simulation scheme described above is the probability $Prob(\mathcal{Y} \mid \mathcal{P})$.*

This observation is due to the independence of the simulated matrix elements which in turn is implied by the independence of the random numbers r. This observation says that the proposed simulation scheme produces every response matrix $\mathcal{Y}$ with probability equaling the probability $\mathfrak{p}(\mathcal{Y})$ (Equation 2.42). In other words, the simulation method creates an approximation of $\mathfrak{p}$. That is why the use of this simulation scheme is well suited for the proposed fit analysis.

The following theorem reveals the relationship of the several expected values appearing in the simulation.

Theorem 1 *Assume that there exists a matrix valued random variable $\mathcal{Y} = (y_{ij})$ $(1 \leq i \leq L,\ 1 \leq j \leq N)$ with the property that the expected value of the ij-th element equals the Rasch conditional probability:*

$$\mathcal{E}(y_{ij}) = P_{ij}. \tag{3.8}$$

Also, assume that $\mathcal{Y}'$ is a random variable with the same properties as that of $\mathcal{Y}$, and that $\mathcal{Y}$ and $\mathcal{Y}'$ are independent.

Then one has the following equalities:

$$\mathcal{E}\left(D^2_{out}(\mathcal{X},\mathcal{Y})\right) - \mathrm{MNSQ}_{out}(\mathcal{X}) = 1, \tag{3.9}$$

$$\mathcal{E}\left(D^2_{in}(\mathcal{X},\mathcal{Y})\right) - \mathrm{MNSQ}_{in}(\mathcal{X}) = 1, \tag{3.10}$$

$$\mathcal{E}\left(D^2_{out}(P,\mathcal{Y})\right) = 1, \tag{3.11}$$

$$\mathcal{E}\left(D^2_{in}(P,\mathcal{Y})\right) = 1, \tag{3.12}$$

$$\mathcal{E}\left(D^2_{out}(\mathcal{Y},\mathcal{Y}')\right) = 2, \tag{3.13}$$

$$\mathcal{E}\left(D^2_{in}(\mathcal{Y},\mathcal{Y}')\right) = 2. \tag{3.14}$$

Proof: First rewrite MNSQ_{out}:

$$NL \cdot \mathrm{MNSQ}_{out} = \sum_{i=1}^{L}\sum_{j=1}^{N}\frac{x_{ij}^2}{P_{ij}Q_{ij}} - 2\sum_{i=1}^{L}\sum_{j=1}^{N}\frac{x_{ij}}{Q_{ij}} + \sum_{i=1}^{L}\sum_{j=1}^{N}\frac{P_{ij}}{Q_{ij}}. \tag{3.15}$$

Then $\mathcal{E}\left(D^2_{out}(\mathcal{X},\mathcal{Y})\right)$ is expressed as:

$$\begin{aligned} NL \cdot \mathcal{E}\left(D^2_{out}(\mathcal{X},\mathcal{Y})\right) &= \mathcal{E}\left(\sum_{i=1}^{L}\sum_{j=1}^{N}\frac{x_{ij}^2}{P_{ij}Q_{ij}}\right) - 2\mathcal{E}\left(\sum_{i=1}^{L}\sum_{j=1}^{N}\frac{x_{ij}y_{ij}}{P_{ij}Q_{ij}}\right) \\ &\quad + \mathcal{E}\left(\sum_{i=1}^{L}\sum_{j=1}^{N}\frac{(y_{ij})^2}{P_{ij}Q_{ij}}\right) \\ &= \sum_{i=1}^{L}\sum_{j=1}^{N}\frac{x_{ij}^2}{P_{ij}Q_{ij}} - 2\sum_{i=1}^{L}\sum_{j=1}^{N}\frac{x_{ij}\mathcal{E}(y_{ij})}{P_{ij}Q_{ij}} + \sum_{i=1}^{L}\sum_{j=1}^{N}\frac{\mathcal{E}(y_{ij})}{P_{ij}Q_{ij}} \\ &= \sum_{i=1}^{L}\sum_{j=1}^{N}\frac{x_{ij}^2}{P_{ij}Q_{ij}} - 2\sum_{i=1}^{L}\sum_{j=1}^{N}\frac{x_{ij}}{Q_{ij}} + \sum_{i=1}^{L}\sum_{j=1}^{N}\frac{1}{Q_{ij}}. \end{aligned}$$

Next, recognize that $(y_{ij})^2 = y_{ij}$ (since y_{ij} only assume values 1 or 0). In the last step, note the assumption that $\mathcal{E}(y_{ij}) = P_{ij}$.

For the difference one has

$$\begin{aligned}
\mathcal{E}\left(D^2_{out}(\mathcal{X},\mathcal{Y})\right) - \mathrm{MNSQ}_{out}(\mathcal{X}) &= \frac{1}{NL}\sum_{i=1}^{L}\sum_{j=1}^{N}\frac{1}{Q_{ij}} - \frac{1}{NL}\sum_{i=1}^{L}\sum_{j=1}^{N}\frac{P_{ij}}{Q_{ij}} \\
&= \frac{1}{NL}\sum_{i=1}^{L}\sum_{j=1}^{N}\frac{1-P_{ij}}{Q_{ij}} \\
&= \frac{1}{NL}\sum_{i=1}^{L}\sum_{j=1}^{N}1 \\
&= \frac{NL}{NL} = 1.
\end{aligned}$$

For the infit index use the weight

$$w_{in} = \frac{1}{\sum_{i=1}^{L}\sum_{j=1}^{N} P_{ij}Q_{ij}} \tag{3.16}$$

Then

$$\mathrm{MNSQ}_{in} = w_{in}\cdot\left(\sum_{i=1}^{L}\sum_{j=1}^{N}x_{ij}^2 - 2\sum_{i=1}^{L}\sum_{j=1}^{N}x_{ij}P_{ij} + \sum_{i=1}^{L}\sum_{j=1}^{N}P_{ij}^2\right). \tag{3.17}$$

On the other side,

$$\mathcal{E}\left(D^2_{in}(\mathcal{X},\mathcal{Y})\right) = w_{in}\cdot\mathcal{E}\left(\sum_{i=1}^{L}\sum_{j=1}^{N}x_{ij}^2 - 2\sum_{i=1}^{L}\sum_{j=1}^{N}x_{ij}y_{ij} + \sum_{i=1}^{L}\sum_{j=1}^{N}y_{ij}^2\right) \tag{3.18}$$

$$= w_{in}\cdot\left(\sum_{i=1}^{L}\sum_{j=1}^{N}x_{ij}^2 - 2\sum_{i=1}^{L}\sum_{j=1}^{N}x_{ij}P_{ij} + \sum_{i=1}^{L}\sum_{j=1}^{N}P_{ij}\right). \tag{3.19}$$

Then write

$$\mathcal{E}\left(D_{in}^2(\mathcal{X},\mathcal{Y})\right) - \mathrm{MNSQ}_{in}(\mathcal{X}) = w_{in} \cdot \sum_{i=1}^{L}\sum_{j=1}^{N} P_{ij}Q_{ij} \tag{3.20}$$

$$= \frac{w_{in}}{w_{in}} = 1. \tag{3.21}$$

This section is concluded by commenting on the Monte Carlo method suggested by Ponoczny (2001).

There, for the simulation of response matrices a similar but in many respect different approach was chosen in an attempt to produce a non-parametric fit test. Ponoczny chose the simulated matrices $\mathcal{Y}$ such that they have the same row marginal sums as $\mathcal{X}$. This raises several issues:

- There have been several studies devoted to the problem of sampling the collection of matrices with given row marginal sums (Snijders, 1991; Rao et. al., 1996; Roberts, 2000; Ponoczny, 2001); and it is also argued by the same authors that the problem in general, due to the size of this set, is unsolvable from a theoretical point of view.

- By fixing the row marginals the process is inherently restricted to the Rasch model, as it is well known that the estimates in the other logistic models are sensitive to the inner structure of the rows of the response matrix. Even though there are estimation methods developed for other IRT models that use only the marginal sums, these are all approximate methods and use some parametric assumption about the underlying data, which is exactly what the Monte Carlo method wants to avoid.

- From the perspective of the probability measure (Equation 2.45) the constant row marginal sums may introduce an unwanted bias to the fit analysis. To make this point more transparent, let C denote the set of response matrices with the same row marginal sums as $\mathcal{X}$. For Ponoczny's method to be unbiased what has to be shown is that

$$\frac{\mathfrak{p}(\mathcal{B} \cap C)}{\mathfrak{p}(C)} = \mathfrak{p}(\mathcal{B}), \tag{3.22}$$

where $\mathcal{B}$ is any set introduced in conjunction with the fit analysis as in Equations 2.44 and 2.46. The rather strong independence condition in Equation 3.22 is hard to believe to hold in general (or for even special cases).

All these problems disappear at once if one chooses the method suggested here.

3.4 Algorithm Defining the New Tests

Following, a family of fit tests is presented. First, a long list of procedures is given and then it will be shown how to combine them to obtain particular members of this family. One introduces the general notation D^2 for the weighted distances introduced in the previous section.

(a) Run a joint maximum likelihood estimation (Wright & Stone (1979) p. 62; Baker (1992) p. 144, also in Chapter 2.7) (or other estimation) procedure on the response matrix $\mathcal{X}$ to obtain the ability and difficulty estimates;

(*b*) Generate K pairs of response matrices $\mathcal{Y}_1^k,\ \mathcal{Y}_2^k\ \ (1 \leq k \leq K)$ using the Rasch model (for precise meaning see the section on the simulation method). The elements of the matrix $\mathcal{Y}_l^k$ are denoted by $y_{l,ji}^k$ $(l = 1, 2)$;

(*c*) Calculate all K D^2 between $\mathcal{X}$ and $\mathcal{Y}_1^k$ according to the following formulae:

$$D^2_{k,out;\mathcal{X}} := D^2_{out}(\mathcal{X}, \mathcal{Y}_1^k) := \frac{1}{NL} \sum_{i=1}^{L} \sum_{j=1}^{N} \frac{\left(x_{ji} - y_{1,ji}^k\right)^2}{P_{ji} Q_{ji}}, \tag{3.23}$$

$$D^2_{k,in;\mathcal{X}} := D^2_{in}(\mathcal{X}, \mathcal{Y}_1^k) := \frac{\sum_{i=1}^{L} \sum_{j=1}^{N} \left(x_{ji} - y_{1,ji}^k\right)^2}{\sum_{i=1}^{L} \sum_{j=1}^{N} P_{ji} Q_{ji}}. \tag{3.24}$$

This provides distributions $D^2_{k,out;\mathcal{X}}$ and $D^2_{k,in;\mathcal{X}}$.

(*d*) Calculate all K D^2 between $\mathcal{P}$ and $\mathcal{Y}_2^k$ according to the following formulae:

$$D^2_{k,out;\mathcal{P}} := D^2_{out}(\mathcal{P}, \mathcal{Y}_2^k), \tag{3.25}$$

$$D^2_{k,in;\mathcal{P}} := D^2_{in}(\mathcal{P}, \mathcal{Y}_2^k). \tag{3.26}$$

This provides distributions $D^2_{k,out;\mathcal{P}}$ and $D^2_{k,in;\mathcal{P}}$.

(*e*) Calculate K D^2 within the set $(\mathcal{Y}_2^k)_{k=1}^K$ defining the distributions $D^2_{k,out;\mathcal{Y}}$ and $D^2_{k,in;\mathcal{Y}}$ by

$$D^2_{k,out;\mathcal{Y}} := D^2_{out}(\mathcal{Y}_2^{k'}, \mathcal{Y}_2^{k''}), \tag{3.27}$$

$$D^2_{k,in;\mathcal{Y}} := D^2_{in}(\mathcal{Y}_2^{k'}, \mathcal{Y}_2^{k''}), \tag{3.28}$$

where (k', k'') are randomly chosen pairs.

Consider the following hypothesis tests (for both outfit and infit; for simplicity the subscripts *out* and *in* are dropped):

(f) H_o: The samples $(D^2_{k;\mathcal{X}} - 1)_{k=1}^K$ and $(D^2_{k;\mathcal{P}})_{k=1}^K$ are from the same population,

H_1: The samples $(D^2_{k;\mathcal{X}} - 1)_{k=1}^K$ and $(D^2_{k;\mathcal{P}})_{k=1}^K$ are NOT from the same population.

(g) H_o : The samples $(D^2_{k;\mathcal{X}})_{k=1}^K$ and $(D^2_{k-\mathcal{Y}})_{k=1}^K$ are from the same population,

H_1 : The samples $(D^2_{k;\mathcal{X}})_{k=1}^K$ and $(D^2_{k-\mathcal{Y}})_{k=1}^K$ are NOT from the same population.

(h) H_o : The sample $(D^2_{k;\mathcal{P}})_{k=1}^K$ and MNSQ are from the same population,

H_1 : The sample $(D^2_{k;\mathcal{P}})_{k=1}^K$ and MNSQ are NOT from the same population.

(i) H_o : The sample $(D^2_{k;\mathcal{Y}} - 1)_{k=1}^K$ and MNSQ are from the same population,

H_1 : The sample $(D^2_{k;\mathcal{Y}} - 1)_{k=1}^K$ and MNSQ are NOT from the same populations.

Note: The two simulated sets are used in step (g) where they are needed to ensure that the two distributions $(D^2_{k;\mathcal{X}})_{k=1}^K$ and $(D^2_{k;\mathcal{Y}})_{k=1}^K$ are independent. To make all these tests independent of one another, one should choose more sets of simulated matrices, although it is avoided here to ease the presentation.

The following four fit tests are developed:

$$MC_{\mathcal{P}\mathcal{X}} \quad : \quad (a) \to (b) \to (c) \to (d) \to (f), \tag{3.29}$$

$$MC_{\mathcal{Y}\mathcal{X}} \quad : \quad (a) \to (b) \to (c) \to (e) \to (g), \tag{3.30}$$

$$MC_{\mathcal{P}M} \quad : \quad (a) \to (b) \to (d) \to (h), \tag{3.31}$$

$$MC_{\mathcal{Y}M} \quad : \quad (a) \to (b) \to (e) \to (i). \tag{3.32}$$

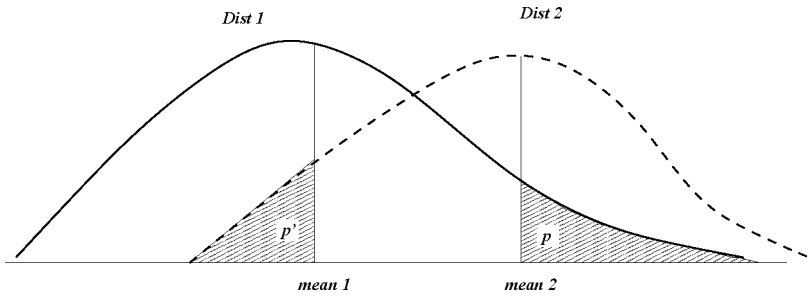

Figure 3.4: Two p-values from two distributions

To actually perform the hypothesis test the method of computing the one-tail probability is chosen by counting the elements in one distribution that are higher (or lower, as the case maybe) than the mean of the other distribution. To be more precise, one defines the p-value for a test by the following:

$$p := \frac{\#\text{ of response matrices with worse fit}}{\#\text{ of all simulated response matrices}}. \tag{3.33}$$

This results in two p-values (denoted by p and p') (as in $MC_{\mathcal{P}\mathcal{X}}$ and $MC_{\mathcal{Y}\mathcal{X}}$) in the case of two distributions (Figure 3.4), and one p-value in the case of a single distribution (Figure 3.5) (as in $MC_{\mathcal{P}M}$ and $MC_{\mathcal{Y}M}$).

Finding a p-value that is larger than the predefined α shows that the data fit the Rasch model, whereas misfit is indicated by a p-value that is less than α. Alternatively, one may report the p-value and let the practitioner to decide on fit.

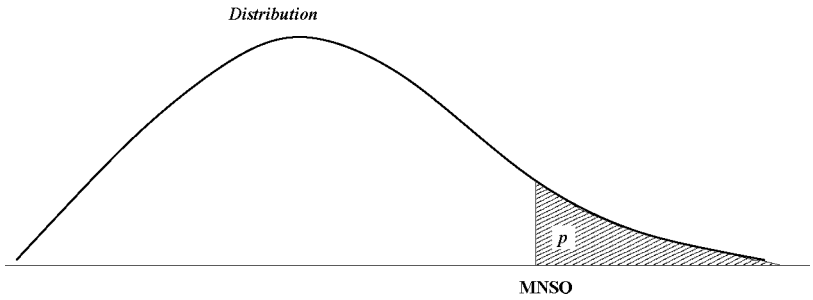

Figure 3.5: p-value from one distribution with MNSQ

3.5 Description and Specifics of the Tests

As a consequence of the innovative simulation technique four tests were created for testing model fit. They have been designed to make use of the simulated distributions extensively. All of them have a sampling distribution simulated, therefore they avoid using assumptions in terms of an existing generic distribution.

The reason for having four tests can be understood by the following. The simulation results in a set of possible response matrices denoted by $\mathcal{Y}$. Using the weighted distance formulae one may construct three distributions of distances:

(a) Distances among $\mathcal{P}$ and $\mathcal{Y}$. This is related to the model which in turn is represented by $\mathcal{P}$. The expected value of this distribution is 1.

(b) Distances among $\mathcal{X}$ and $\mathcal{Y}$. This is related to the data represented by $\mathcal{X}$. The expected value is MNSQ+1 (see Theorem 1).

(c) Distances among $\mathcal{Y}$. Here again, only the model is used (implicitly, since $\mathcal{Y}$ is generated using $\mathcal{P}$). The expected value of this distribution is 2.

The goal of fit analysis is to compare the model and the data ($\mathcal{P}$ and $\mathcal{X}$). There are four ways to do this.

- $MC_{\mathcal{P}\mathcal{X}}$: Compare (a) and (b),
- $MC_{\mathcal{Y}\mathcal{X}}$: Compare (b) and (c),
- $MC_{\mathcal{P}M}$: Compare (a) to MNSQ,
- $MC_{\mathcal{Y}M}$: Compare (c) to MNSQ.

There are other comparisons possible which are left out. One may compare MNSQ to (b) but that would not check the fit of the model, rather it would constitute a check for the simulation. The same can be said about the comparison of (a) and (c) as this lacks explicit reference to the data $\mathcal{X}$ it would only check how accurate the simulation is. A more sound test for the appropriateness of the sample $\mathcal{Y}$ will be outlined in the stability study presented in Chapter 4.

In the previous section the algorithms for developing all four tests have been presented. Now, each test will have an in-depth discussion about its properties, rationale, and its underlying geometric pictures will be presented.

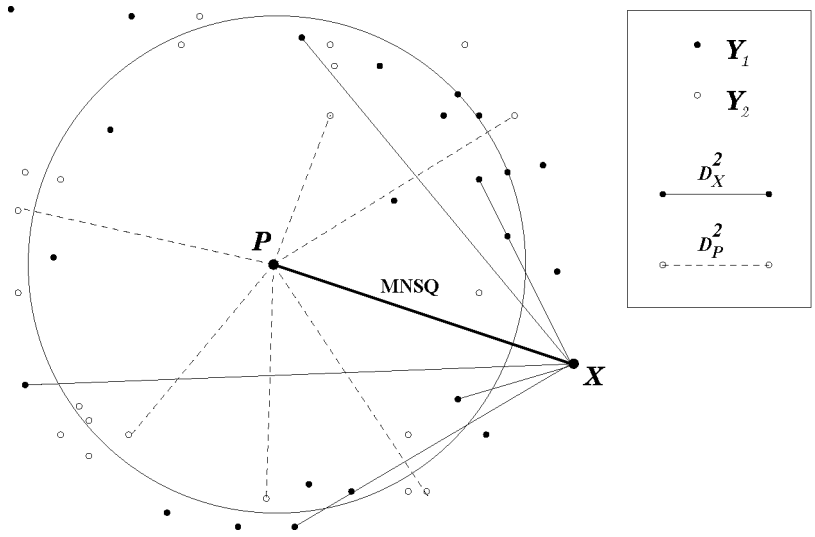

Figure 3.6: Geometry Behind $MC_{\mathcal{PX}}$

3.5.1 $\mathbf{MC}_{\mathcal{PX}}$

The first test aims to find out whether there is a significant difference between the empirical distribution built from $\mathcal{X}$ and $\mathcal{Y}^k$ and the sampling distribution generated using $\mathcal{P}$ and $\mathcal{Y}^k$.

More precisely, the distribution of $D^2_{k;\mathcal{X}}$, defined in Equations 3.23 and 3.24, represents an empirical distribution furnished by the squared differences of the response matrix $\mathcal{X}$ and the generated matrices $\mathcal{Y}^k_l$. The second distribution is made out of $D^2_{k;\mathcal{P}}$, defined in Equations 3.25 and 3.26, which represents the squared distances between $\mathcal{Y}^k_2$ and the $\mathcal{P}$ matrix (Figure 3.6). Then, a significance test is need to be performed to find the relevant p-values. This is simply done by counting the portion

of cases beyond the intersection of the mean and the tail of the other distribution depending on which tail probability is needed (Figure 3.4).

The motivation behind this test is as follows: MNSQ fit analysis aims to find and assess the distance between $\mathcal{P}$ and $\mathcal{X}$. Good fit is obtained if that distance is close to 1. To make the comparison, the test "blows up" $\mathcal{P}$ and substitutes it with a set of response matrices $\mathcal{Y}$. If $\mathcal{P}$ is close to $\mathcal{X}$ then the average distance between $\mathcal{P}$ and the set of $\mathcal{Y}$s is close to the average distance of $\mathcal{X}$ and $\mathcal{Y}$s. Unlike the traditional MNSQ fit tests, this time one not only has the average values but both distributions are present making a realistic comparison possible.

An awkward feature of this test is that 1 has to be subtracted from the later distribution. This 1 is the expected value of the distance between $\mathcal{P}$ and $\mathcal{X}$ (for both outfit and infit) which is represented by a circle in Figures 3.6 to 3.9. This topic was detailed previously in Theorem 1.

3.5.2 $\mathbf{MC}_{\mathcal{Y}\mathcal{X}}$

The test $MC_{\mathcal{Y}\mathcal{X}}$ aims to serve as an alternative to $MC_{\mathcal{P}\mathcal{X}}$. It also requires two distributions similarly to the previous test. Although, the empirical distribution $(D^2_{k,\mathcal{X}})$ stays the same the nature of the sampling distribution has changed. Note that the two simulated sets ensure that the two distributions $(D^2_{k;\mathcal{X}})_{k=1}^K$ and $(D^2_{k;\mathcal{Y}})_{k=1}^K$ are independent (Figure 3.7).

If the model fits the data, then $\mathcal{X}$ is indistinguishable from the set of model generated response matrices $\mathcal{Y}$. That is, one should find that the average distance

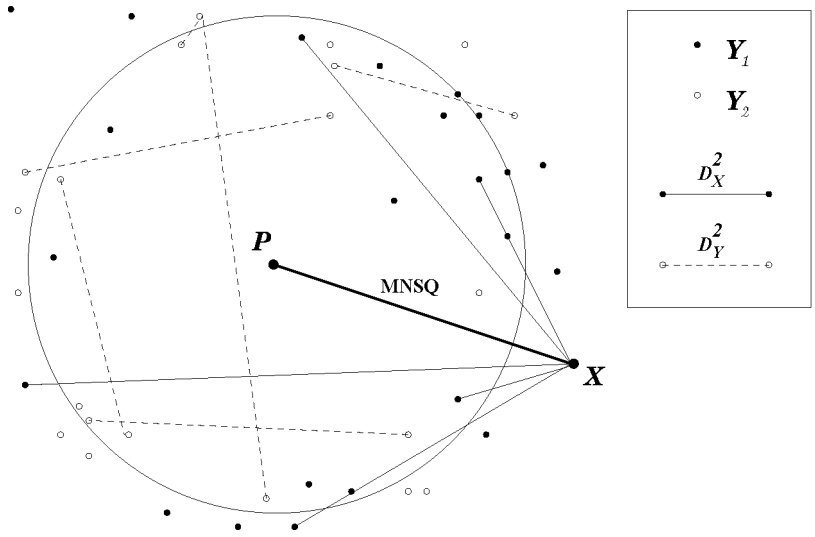

Figure 3.7: Geometry Behind $MC_{\mathcal{Y}\mathcal{X}}$

between $\mathcal{X}$ and the set of $\mathcal{Y}$ is the same as the average distance among the $\mathcal{Y}$ matrices. Here, again, one has two distributions resulting in two p-values (p and p').

3.5.3 $\mathbf{MC}_{\mathcal{P}M}$

This test reveals whether there is a significant difference between the sampling or modeled distribution generated with $\mathcal{P}$ and $\mathcal{Y}^k$, and MNSQ statistics.

$MC_{\mathcal{P}M}$ requires only one distribution and one fit statistic. The distribution is constructed from $D^2_{k;\mathcal{P}}$, defined in Equations 3.25 and 3.26 the squared residuals between $\mathcal{Y}^k_2$ and the P matrix, just like in the first test (Figure 3.8). This test might resemble the test developed by Wright and Stone (1979), however, there is a

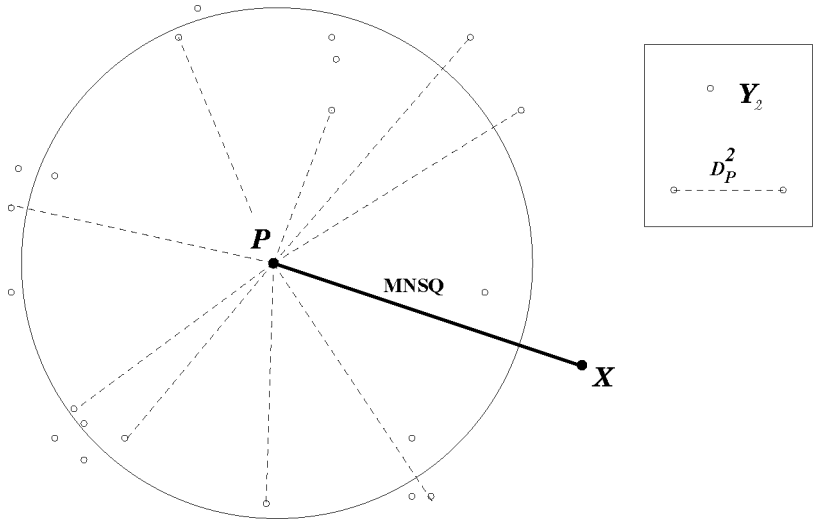

Figure 3.8: Geometry Behind $MC_{\mathcal{P}M}$

significant difference between the two methods. In this case, no assumption is made regarding the sampling distribution; rather it is generated "on the fly" based on the model and response matrix characteristics. The test is distribution-free, hence the p-value is not influenced by the violation of assumptions and will give unbiased results. Issues related to distributional assumptions in Rasch fit analysis have been discussed in the previous section.

The rationale behind this test is that any simulated $\mathcal{Y}$ matrix could be a perfectly acceptable response matrix. As always, the set of response matrices, $\mathcal{Y}$s, are created to represent the model. The distance MNSQ is the traditional residual-based fit statistics. This time, however, one does not have to evaluate it against an artificial

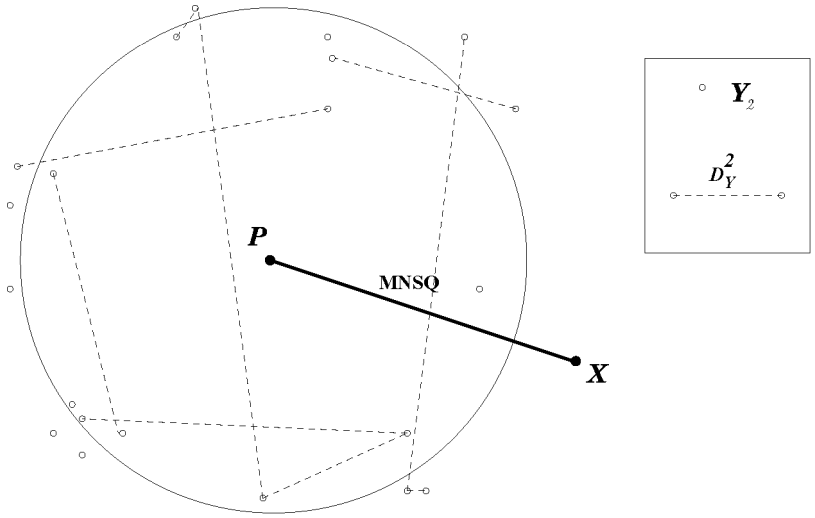

Figure 3.9: Geometry Behind MC_{y_M}

(and faulty) normality assumption. Rather, the Monte Carlo sampling distribution is created by the computer and the p-value is calculated on his generated distribution (Figure 3.5). The idea is very natural and one believes that it has not yet become common practice only because robust computers are just becoming available, and without them the Monte Carlo fit test could run for days.

3.5.4 $\mathbf{MC}_{y_M}$

Yet another extension of the ideas above is the test denoted by MC_{y_M}. The underlying geometric picture is explained as follows (Figure 3.9). To assess the distance between $\mathcal{X}$ and $\mathcal{P}$ first one "blows-up" $\mathcal{P}$ into a set of response matrices and then

substitutes $\mathcal{X}$ by another set of response matrices. The average distance between these two sets is then evaluated against MNSQ. If the model fits, then $\mathcal{X}$ is equivalent with the generated set of $\mathcal{Y}$s, so MNSQ should be the same as the average. Again, a closer look at the structure of the test reveals that 1 has to be subtracted from the distribution before comparison. Having only one distribution and a fixed number (MNSQ) this test provides only one p-value (Figure 3.5).

3.6 Case of General Fit Indices

In this section the author indicates how the above described procedure can be extended to other IRT models and to other indices.

Let us accept for the purposes of this discussion that an IRT model is fully represented by its response probability matrix $\mathcal{P}$. As before, the elements of $\mathcal{P}$ are the probabilities of the actual performance levels given all the model parameters. Any IRT model defines the set of response matrices, that is the set of matrices with size $N \times L$ and with entries according to the model. For a partial credit model, for example, the entries are restricted to the range of possible response levels.

The simulation method should be changed in accordance with the model. The actual method varies from theory to theory but the content is always the same: one has to sample the set of response matrices so that the relative frequency of a matrix equals its probability calculated from $\mathcal{P}$. The crucial difference between models lies in the way one may find the estimates of the parameters, but that is no concern of the Monte Carlo method outlined here.

According to Drasgow, Levine, & McLaughlin (1987) and Klauer (1995) and Snijders (2001) almost all fit statistics can be expressed in the general form

$$V = \sum_{i=1}^{L} \sum_{j=1}^{N} (x_{ji} - P_{ji}) w_{ji} \tag{3.34}$$

or

$$V^* = \sum_{i=1}^{L} \sum_{j=1}^{N} (x_{ji} - P_{ji})^2 v_{ji}, \tag{3.35}$$

where w_{ji} and v_{ji} are appropriate weight functions. As can be shown, the procedure in the previous section generalizes in a straightforward manner to any fit statistics given by Equations 3.34 and 3.35. To be more specific, let us briefly indicate how one can formulate this procedure in the presence of a general fit index V and V^* (Equations 3.34 and 3.35). For simplicity only $MC_{\mathcal{P}V}$ the index corresponding to $MC_{\mathcal{P}M}$ is covered.

Let us again start with the assumption that the parameter estimates have already been obtained. Then, using a set of simulated matrices $\mathcal{Y}^k,\ k = 1, \ldots, K$ as before one creates distributions

$$V_k := V(\mathcal{P}, \mathcal{Y}) \ := \ \sum_{i=1}^{L} \sum_{j=1}^{N} (y_{ji}^k - P_{ji}) w_{ji}, \tag{3.36}$$

$$V_k^* := V^*(\mathcal{P}, \mathcal{Y}) \ := \ \sum_{i=1}^{L} \sum_{j=1}^{N} (y_{ji}^k - P_{ji})^2 v_{ji}. \tag{3.37}$$

Then, one performs the hypotheses testing to see if

$$V = V(\mathcal{P}, \mathcal{X}) \quad \text{and} \quad V^* = V^*(\mathcal{P}, \mathcal{X}) \tag{3.38}$$

belong to the same population as the samples (V_k) and (V_k^*), respectively.

It is now easy to understand how one may extend the Monte Carlo procedure to a much wider context. For *any* index calculated from $\mathcal{X}$ and $\mathcal{P}$ one may produce a simulated set of response matrices and calculate the same index for $\mathcal{Y}$ now in place of $\mathcal{X}$. The observation made in Chapter 3.3 about the correspondence between the relative frequency of $\mathcal{Y}$ and the probability $\mathfrak{p}(\mathcal{Y})$ of $\mathcal{Y}$ makes it possible to find the p-value of the index in question.

Credible tests must possess good psychometric characteristics in order to be used with trust. Thus, after introducing the new Monte Carlo fit tests the next step to proceed to is to investigate the validity and stability of the new tests.

CHAPTER 4

ANALYSES AND RESULTS

4.1 Overview

The rationale behind the new testing method is the creation of a sampling distribution representing a model behavior and the creation of a second distribution that shows the characteristics of observed data. In distribution generation the main idea is to use the distance interpretation of MNSQ. From this point of view, the goal of the fit test is to find the distance between $\mathcal{P}$ and $\mathcal{X}$. There are four ways to test model fit employing simulated distributions. The process of executing these four tests was introduced in the previous chapter. Now, the investigation extends to "how good these tests are", "how they relate to the old MNSQ test", and "what their behavior is under different conditions"? This chapter includes two major sections that deal with the validity and stability of the new tests.

A validity study is presented to contrast the new method with the old MNSQ test and to highlight the fundamental differences. Within this framework a complex study explores the association between the $MC_{\mathcal{P}M}$ Monte Carlo test and the traditionally used MNSQ outfit test utilizing different kinds of response matrices of different sizes. In order to make the comparison on a global scale 8×8, 25×25, 50×50, and

100×100 response matrices will be generated and fit indices and their appropriate p-values calculated to determine if there is any functional relationship between the Monte Carlo (MC) and the traditional MNSQ outfit tests. Due to the time consuming nature of this study the focus is narrowed down to the outfit version of $MC_{\mathcal{P}M}$. For the purpose of the validity study real test data will be used from Mathematics 116 midterm and final examinations. These exams were given to undergraduate students by the Department of Mathematics at the Ohio State University in Spring, 2002. The midterm exam contained 24 multiple choice questions (with no partial credits) and was administered to 82 students. The final exam contained 48 multiple choice questions (with no partial credits) and was given to 82 students. Both pools of examinees were split into two approximately equal subgroups. The following four data matrices were derived: 82×48, 40×48, 40×24, and 82×24.

A related study in this section investigates the influence of the original score distribution on the outcome of $MC_{\mathcal{P}M}$ test. The new MC tests employ the response matrix $\mathcal{X}$ in their initial step to obtain parameter estimates. In an extreme case, when a large number of spuriously high test scores are obtained from the pool of examinees will result in a negatively skewed score distribution or, spuriously low test scores will yield a positively skewed distribution. None of these cases are atypical in educational measurement. Since the fit index depends highly on the inner structure of the response matrix (not only on the marginal sums) one does not expect any simple correspondence between the distribution of marginal sums and the actual fit value. With the exact same marginal sums a matrix can be very well fitting (a response matrix in which rows are close to pure Guttmann patterns is an example). However,

the fit can be made worse by introducing several aberrant responses and keeping the marginal sums intact. Data for this study was obtained from a licensing exam of the International Hearing Society.

The last section investigates the stability of the MC p-values with respect to the number of simulated response matrices (K). The goal of this study is to find an empirical threshold for K, a number K_t such that for all $K > K_t$ the fit analysis gives essentially the same result. This threshold is also required to be as small as possible, since the computational time increases with K.

4.2 Validity Study

4.2.1 First Look at the MC Statistics

This analysis presents the four new Monte Carlo (MC) fit tests and the old MNSQ tests, form both the infit and the outfit point of view. (One actually has six p-values because there are 2 tests that have two p-values, as explained earlier). The goal of this analysis is to present the proposed MC p-values along with their test statistics. By test statistics here one understands the D^2 values defined in Equations 3.23 through 3.28. The concrete values of these statistics are used in defining the MC p-values and they also provide some useful insight about the MC fit analysis.

Exam data were obtained from a Mathematics 116 final examination (with 82 subjects and 48 items). MNSQ outfit and infit indices were calculated using Equations 2.32 and 2.33 and then converted, using the Wilson-Hilferty transform (Equations 2.34 and 2.35) to an approximate normal distribution $N(0,1)$ for easy evaluation of

p-value. These values are reported under p_{χ^2} in Table 4.1, which also shows the result of the computations of D^2 statistics and the corresponding standard deviations.

In Table 4.1 the column of $p_{*\mathcal{X}}$ shows the p-values for $MC_{\mathcal{Y}\mathcal{X}}$ and $MC_{\mathcal{P}\mathcal{X}}$ (for both infit and outfit). For example, the p-value for $MC_{out,\mathcal{Y}\mathcal{X}} = 0.229$. The columns of p_{*M} and $p'_{*\mathcal{X}}$ are explained similarly. For this run the number of simulated matrices chosen is the ample $K = 10,000$.

Statistic	**Value**	**Std.Dev.**	p_{χ^2}	$p_{*\mathcal{X}}$	$p'_{*\mathcal{X}}$	p_{*M}
MNSQ_{out}	0.9475		0.009			
$D^2_{out;\mathcal{X}}$	0.9464*	0.0806				
$D^2_{out;\mathcal{Y}}$	0.9991*	0.1093		0.229	0.323	0.328
$D^2_{out;\mathcal{P}}$	1.0002	0.0747		0.225	0.234	0.240
MNSQ_{in}	0.9832		0.444			
$D^2_{in;\mathcal{X}}$	0.9832*	0.0444				
$D^2_{in;\mathcal{Y}}$	0.9996*	0.0498		0.356	0.378	0.378
$D^2_{in;\mathcal{P}}$	0.9998	0.0230		0.356	0.238	0.238

Table 4.1: MNSQ and D^2 statistics and corresponding p-values. (Values with * are reduced by one.)

The MNSQ outfit statistic is 0.9475 which represents the traditionally used statistic calculated by Equation 2.32. Its p-value is 0.009 - a significant misfit.

$D^2_{out-\mathcal{X}}$ has been defined in Equation 3.23. It is the mean of the squared distances among $\mathcal{X}$ and the simulated $\mathcal{Y}^k$ set. This mean is the estimate of the expected value of $D^2_{out-\mathcal{X}}$. According to Theorem 1 this mean (more precisely, its expected value) is the same as MNSQ after subtracting 1. For easy comparison the table contains this modified value. Indeed, the difference $0.9475-0.9464 = 0.0011$ is close to 0 indicating the appropriateness of the simulation. The standard deviation of $D^2_{out-\mathcal{X}}$ is 0.0806 which shows that the expected value of $D^2_{out-\mathcal{X}}$ and of MNSQ and $D^2_{out-\mathcal{X}}$ is within 1 standard deviation. Recall that, the intention in hypothesis testing concerning model fit is to assess the difference between MNSQ and its expected value 1. Considering this case, the expected value is within one standard deviation from MNSQ. This in itself is not a strong ground for judging model fit, but since the distribution of $D^2_{out-\mathcal{X}}$ is also at our disposal the corresponding p-values can be obtained.

$D^2_{out-\mathcal{Y}}$ is the mean of the squared distances among the elements of the simulated set. Its expected value is 2, that is why the value reported is really $D^2_{out-\mathcal{Y}} - 1$. Note, that its actual value is very close to 1, indicating once more the goodness of the simulation. The standard deviation is 0.1093, so the actual MNSQ value is again inside the interval of 1 standard deviation (of $\mathcal{E}(\text{MNSQ}) = 1$).

$D^2_{out-\mathcal{P}}$ is the mean of the squared distances between $\mathcal{P}$ and the members of the simulated $\mathcal{Y}^k$ set. Here, the expected value is 1, and the obtained 1.0002 reinforces the representativeness of the simulated set. MNSQ falls inside the interval of 1 standard

deviation (0.00747) here, as well. The rest of the table contains the outfit p-values. All of these indices indicate moderate fit, contradicting p_{χ^2}.

For infit one could repeat the entire analysis almost word by word. The only significant difference is that the traditional MNSQ infit shows very good fit, while the MC tests indicate good to moderate fit. It turns out, it is a special feature of infit to be extremely lenient.

The newly introduced MC indices, however, make good sense even with the use of the more lenient infit. The reason is that even though the actual value of infit is lenient so are the infit values for the simulated response matrices. The insensitivity of infit to aberrant responses survives in the fact that the MC infit p-values tend to be higher than their outfit counter parts. An explanation for this might be that the weight of outfit penalizes outlier responses while the constant weight of infit only measures the difference between $\mathcal{P}$ and $\mathcal{X}$, irrespective to how aberrant the pattern may be.

4.2.2 Comparison Study

In this study the goal of the investigation is to see if all MC tests display the same findings in relation with each other. Here, the effort is made to see how consistent the new tests are and whether they contradict the traditional tests. Note, that the possibility of contradiction was indicated in the previous study. For the purposes of this analysis four data sets were used which had been described in the Overview.

The comparison between the traditional MNSQ and the new MC tests can be seen easily (Table 4.2 shows all p-values) as the results are consistently the same for all.

Test	p-value	82×48	40×48	40×24	82×24
$MNSQ_{out,zstd}$	p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**
$MC_{\mathcal{P}\mathcal{X}}$	$p_{\mathcal{P}\mathcal{X}}$	0.2970	0.2194	0.2874	0.2536
$MC_{\mathcal{P}\mathcal{X}}$	$p'_{\mathcal{P}\mathcal{X}}$	0.3174	0.1796	0.3000	0.2358
$MC_{\mathcal{Y}\mathcal{X}}$	$p_{\mathcal{Y}\mathcal{X}}$	0.2986	0.2192	0.2802	0.2536
$MC_{\mathcal{Y}\mathcal{X}}$	$p'_{\mathcal{Y}\mathcal{X}}$	0.3842	0.2644	0.3654	0.3386
$MC_{\mathcal{P}M}$	$p_{\mathcal{P}M}$	0.3200	0.1818	0.2988	0.2338
$MC_{\mathcal{Y}M}$	$p_{\mathcal{Y}M}$	0.3856	0.2660	0.3652	0.3366

Table 4.2: Comparison between MNSQ and MC indices for four different exam data.

The new MC tests rarely show significant misfit (not even at the strict $\alpha = 0.1$ level), while MNSQ outfit shows significant (at $\alpha = 0.05$) misfit 2 out of 4 cases (these are highlighted in the table).

Although, all MC tests employ the same method obtaining its relevant p-value, each has its unique characteristics. Due to their differences the resulting p-values are slightly different. It is not very difficult to understand their differences in the light of the general theory of model fit (given in Chapter 2). Given any new index the set of misfitting matrices may differ. One cannot expect that those matrices which are further away with respect to one distance is going to be further away with respect to another distance (see the geometrical interpretation in Figures 3.6 through 3.9). As

long as they indicate the same message there is no transparent reason to discard any of them.

It is worthwhile to note that the order among the four exams in terms of p-values is exactly the same for all MC tests. The lowest p-value was obtained for exam 40×48 and the highest for 82×48, considering all six MC p-values. Further, even MNSQ seems to follow the same order. This suggests a hidden functional relationship between the new and the old indices.

The next, global analysis study, aims to reveal the association between MNSQ outfit and the MC_{PM} in greater depth than it was presented in this example.

4.3 Global Fit Analysis

This comprehensive study aims to examine the relationship of MC_{PM} test and the old $MNSQ_{outfit}$ test on a large global scale. The observed response matrices $\mathcal{X}$ are obtained through simulation and then entered to the testing process. For thorough investigation four different response matrix sizes were used: 8x8, 25x25, 50x50 and 100x100. For each size 1000 response matrices were generated.

4.3.1 Statement of Problem

In any discrete item response theory model the number of response matrices $\mathfrak{n}$ of fixed size is finite. For any fixed significance level $\alpha \in [0, 1]$ this finite set is divided into two parts: the set $\mathcal{F}_\alpha$ for which a particular item response theory model fits with probability higher than α, and the rest of the set. If the cardinality of $\mathcal{F}_\alpha$ is $\mathfrak{f}_\alpha$ then

the ratio $\frac{f_\alpha}{n}$ expresses the extent to which the model holds in general. The goal of this study is to find this probability.

What makes the task challenging is the fact that the set of response matrices is humongous. Therefore, the first step in the global fit analysis is to develop a method for sampling the set of response matrices. The other observation that makes this step difficult is that a completely random sample has to be extremely large to contain matrices which do not show almost perfect fit. These difficulties are overcome by introducing a matrix generating method dictated by understanding real test situations.

For any response matrix from the sample created above in the next step the Monte Carlo fit analysis is performed. To obtain an estimate of the ratio $\frac{f_\alpha}{n}$ one counts the number of fitting response matrices in the sample and divides this number by the cardinality of the sample.

All the numerical calculations performed in this study made use of the Rasch estimation engine developed for a comprehensive Rasch analysis program (Antal & Antal, 2003). All the programs used here were written using the Perl language (Wall et. al., 2001).

4.3.2 Sampling of $M_{resp}(N, L)$

To draw a sample from the set of response matrices suitable for the global fit analysis is not a straightforward task. The nature of item response theory models shows that for a generic response matrix the model fits. In real life, the response matrices are not generic. This means that a real life dichotomous response matrix has many more 1 entries than 0 entries. That is, in a realistic test situation one assumes that the

subjects usually give correct responses. Note, however, that this assumption is by no means a requirement of the results presented here. It is only used to help us find a good way of generating response matrices for which fit is less expected.

Note, also, that using the Rasch model to generate response matrices (see Equation 3.6) is highly undesirable, as a sample so generated would be biased toward the Rasch model.

A rather large family of response matrix generating methods can be described as follows. First one chooses a matrix $\mathcal{G}$ of the same size as the response matrix one wishes to generate with entries $g_{ji} \in [0,1]$. $\mathcal{G}$ is called the generating matrix. The entry x_{ji} in a simulated response matrix $\mathcal{X}$ is then set by generating a uniformly distributed random number $r \in [0,1]$ and then defining

$$x_{ji} = \begin{cases} 1 \text{ if } r \leq g_{ji}, \\ 0 \text{ otherwise.} \end{cases} \tag{4.1}$$

Choosing $g_{ji} = \frac{1}{2}$ for all (j,i) would be the generic choice mentioned above, unfortunately giving mostly well fitting matrices with respect to *any* fit index. Of course, if the generating process was left to run indefinitely it would produce every response matrix with non zero frequency. One does not have that much time, though.

It is sufficient to obtain a great many "interesting" response matrices by choosing $g_{ji} = t$ for all (j,i), where t is sufficiently close to 1, (or 0 if one desires to model mostly incorrect responses). The acceptable range for t depends on the actual size (N,L) of the response matrix. A little experimentation shows that $t = 0.075$ works very well for $(N,L) = (25,25)$ to $(N,L) = (100,100)$.

There are two effects acting in opposite directions that need to be considered when generating response matrices. On one hand, one wants to generate response matrices which are "appropriate" from the point of view of fit analysis, that is, they are mainly filled with 1 and with only a small number of 0 is added, or vice-versa. This pushes t close to 1 (or to 0). The effect of the proliferation of 1 is that a few "misplaced" 0 can add substantially to the non-fitting direction. "Misplaced" means that for a subject with high ability the incorrect response is given to a relatively easy item.

On the other hand, since for the generating matrix an estimation procedure has to be performed, the matrix has to be without constant rows or columns; in other words, it has to be *"clean"*. Of course, it is impossible to randomly generate a clean matrix - the algorithm should discard un-clean matrices. If there are too many matrices to be discarded the computational time will increase uncomfortably. The closer t is to 1 the more likely it is to simulate an un-clean matrix.

If one desires, one may tune $\mathcal{G}$ to obtain response matrices with prescribed response patterns (several Guttmann like or reverse Guttmann like rows, patterns reflecting time pressure, and so on).

4.3.3 Results of the Global Analysis

The global fit analysis starts with drawing a big enough sample from the set of response matrices keeping in mind the subtleties discussed in the previous section. Next, for each generated response matrix the traditional fit analysis (for comparison purposes) and the Monte Carlo fit analysis are performed.

As previously indicated, using all the newly introduced fit induces, while feasible in principle, could have resulted in a very lengthy computation. The computation with $MC_{\mathcal{P}M}$ for the largest (100×100) size lasted for two days! Since raising the number of simulated matrices (K) in the fit analysis to the more desirable 5000 would have changed the required running time to 10 days. The less stable indices used at the $K = 1000$ level were avoided. Also, the comparison of the new MC indices in the previous section showed that they have some degree of consistency. Since this study is not designed to further this comparison, there seems to be no reason to include all the new MC indices.

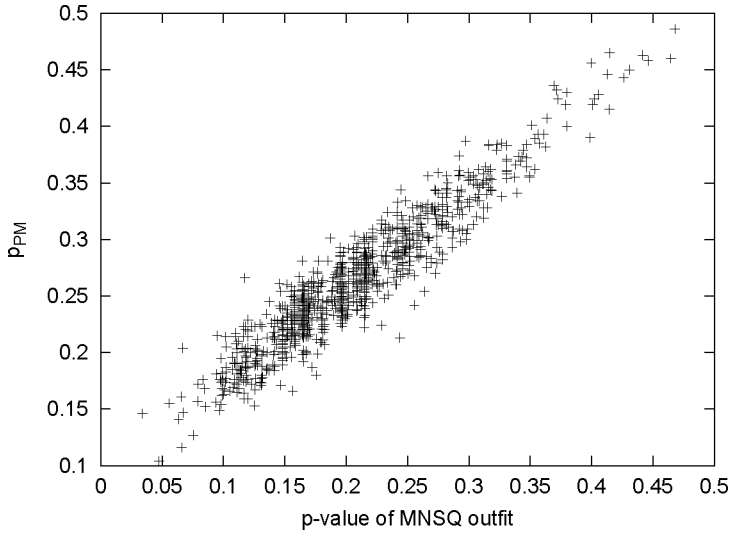

Figure 4.1: p-values of MNSQ outfit vs. $MC_{\mathcal{P}M}$ (8×8).

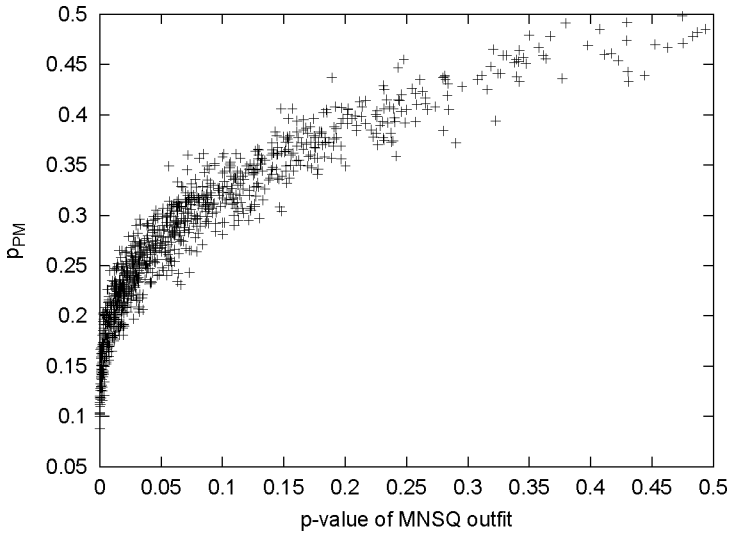

Figure 4.2: p-values of MNSQ outfit vs. $MC_{\mathcal{P}M}$ (25×25).

The first series of figures shows the comparison of MNSQ outfit and $MC_{\mathcal{P}M}$ p-values. The size of the response matrices were 8×8 (Figure 4.1), 25×25 (Figure 4.2), 50×50 (Figure 4.3), and 100×100 (Figure 4.4). For all cases 1000 response matrices were generated and the Monte Carlo fits were calculated with $K = 1000$ simulations.

Several observations can be made about these figures. The first is that the chosen method of sampling response matrices that are problematic from the fit point of view is successful when the size is at least 25×25. For the three larger sizes the majority of the simulated response matrices show significant MNSQ misfit. Moreover, the larger the size the higher the portion of MNSQ misfitting matrices. The smallest (8×8) case requires special attention, since here the simulation method was not successful

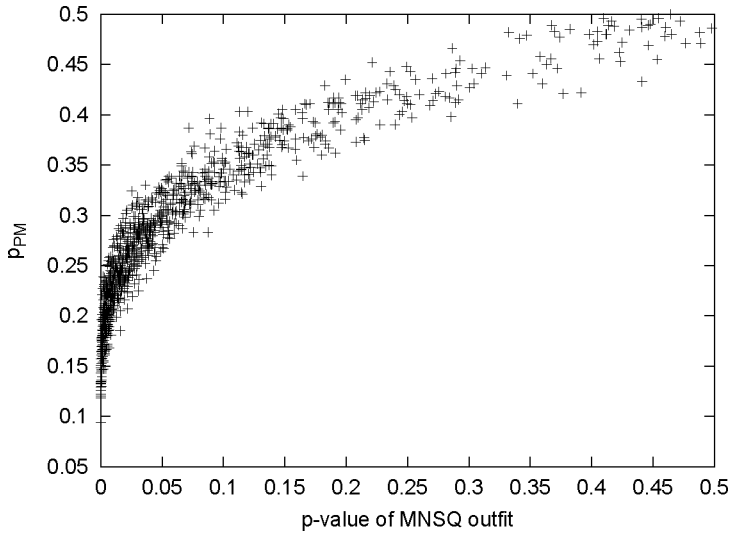

Figure 4.3: p-values of MNSQ outfit vs. $MC_{\mathcal{P}M}$ (50×50).

in selecting misfitting matrices. By pushing the threshold to $t = 0.075$ does not appear to be a good idea, as the number un-clean matrices (those with constant rows or columns) grows rapidly among the simulated response matrices increasing the running time significantly.

More importantly, for all the simulated response matrices the Monte Carlo index shows fit with higher than $\alpha = 0.05$ confidence level, and only a very few display Monte Carlo misfit at the $\alpha = 0.1$ confidence level. Adding, that the Monte Carlo fit index is not just yet another fit index but it is the simulation version of the un-approachable "exact fit index", this result should be read as a very strong conjecture forming the heart of this study (formulated rather loosely here):

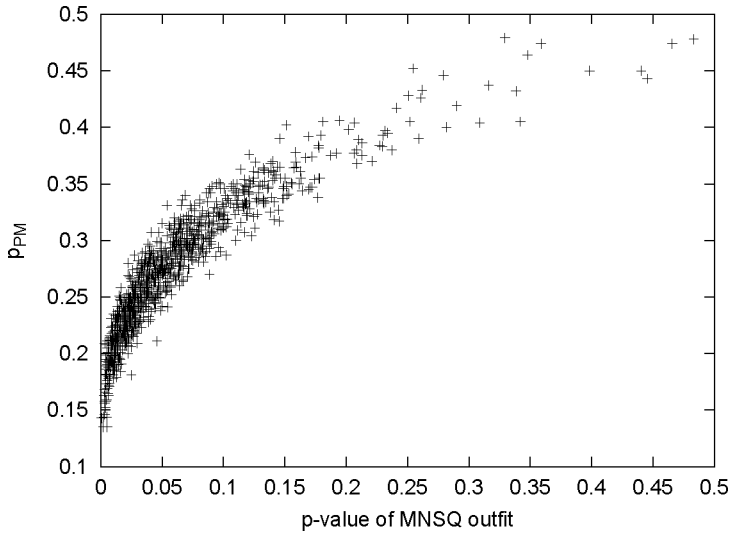

Figure 4.4: p-values of MNSQ outfit vs. $MC_{\mathcal{P}M}$ (100×100).

Conjecture 1 *The Rasch model almost never shows misfit.*

To provide a better understanding of this comparison the frequency distributions of MNSQ and $MC_{\mathcal{P}M}$ is presented in Figures B.1 - B.4 in Appendix. The information conveyed by these figures is again that for larger matrices the MNSQ shows misfit with overwhelming frequency even at the $\alpha = 0.05$ level. For the Monte Carlo index there is not a single matrix showing worse than 5% misfit, and the number of misfitting matrices at the $\alpha = 0.1$ level is at most 1.

Table 4.3 shows the experimental global fit values. The values presented are the relative frequencies of fitting matrices at the level $\alpha = 0.05$. The relative frequencies of fitting matrices at the level $\alpha = 0.1$ are listed in parentheses.

	8×8	25×25	50×50	100×100
MNSQ	0.998 (0.974)	0.493 (0.295)	0.393 (0.278)	0.438 (0.177)
$MC_{\mathcal{P}M}$	1 (1)	1 (0.999)	1 (0.999)	1 (1)

Table 4.3: The experimental global fit indices of MNSQ and $MC_{\mathcal{P}M}$ for four different sizes. The listed values are the probabilities of fit at $\alpha = 0.05$ ($\alpha = 0.1$) significance levels.

4.4 Distributional Considerations

To test the distributional dependence of the new MC indices one might be tempted to use several response matrices with different score distributions and to compare the resulting fit indices. Let us explain why this approach is not tenable.

It is a remarkable feature of the Rasch model, that parameter estimates can be obtained, without any additional assumption, from the marginal sums of the response matrix. In other words, the parameter estimates are not very sensitive to the inner structure of the response matrix. From this point of view there seems to be a small difference between classical test theory and Rasch IRT. The main difference, however, is that the *Rasch fit* analysis is strongly influenced by the structure of the response matrix. One cannot, without punishment, permute the elements, even if the marginals are kept the same. From this, it is clear that there could be no reasonable correspondence between score distributions and fit values. Nevertheless, three examples with different score distributions are provided. All of them are obtained from

Statistic	Value	Std.Dev.	p_{χ^2}	$p_{*\mathcal{X}}$	$p'_{*\mathcal{X}}$	p_{*M}
MNSQ_{out}	0.9919		0.1383			
$D^2_{out;\mathcal{X}}$	0.9912	0.0178				
$D^2_{out;\mathcal{Y}}$	0.9997	0.0221		0.3230	0.3580	0.3680
$D^2_{out;\mathcal{P}}$	1.0004	0.0138		0.3110	0.2380	0.2560
MNSQ_{in}	0.9903		0.4358			
$D^2_{in;\mathcal{X}}$	0.9897	0.0145				
$D^2_{in;\mathcal{Y}}$	0.9994	0.0161		0.2580	0.2610	0.2680
$D^2_{in;\mathcal{P}}$	0.9999	0.0077		0.2450	0.0840	0.1040

Table 4.4: Fit Test for Total IHS exam

an International Licensing Examination administered by the International Hearing Society (IHS), which is a professional organization of hearing instrument dispensers in the USA. They offer this exam to US states and Canadian provinces which in turn provide the license to practice as hearing instrument dispenser in their respective state/province. Their exam is based on a competency model which was derived from role delineation studies conducted in 1981, 1990 and 2000 (D'Costa, 2003).

The first response matrix corresponds to the total exam while the other two were obtained by carefully splitting this exam into two tests. The splitting was performed in a way to result in two sets with different distributional properties. Score distributions are shown in Figures A.1, A.3, and A.2 (in the Appendix). The results of the fit analyses are presented in Tables 4.4, 4.5, and 4.6.

Statistic	**Value**	**Std.Dev.**	p_{χ^2}	$p_{*\mathcal{X}}$	$p'_{*\mathcal{X}}$	p_{*M}
MNSQ_{out}	1.0032		0.3830			
$D^2_{out;\mathcal{X}}$	1.0040	0.0253				
$D^2_{out;\mathcal{Y}}$	1.0011	0.0319		0.4540	0.4610	0.4730
$D^2_{out;\mathcal{P}}$	1.0011	0.0200		0.4540	0.4310	0.4450
MNSQ_{in}	0.9939		0.4494			
$D^2_{in;\mathcal{X}}$	0.9942	0.0189				
$D^2_{in;\mathcal{Y}}$	1.0004	0.0206		0.3580	0.3750	0.3750
$D^2_{in;\mathcal{P}}$	1.0006	0.0080		0.3580	0.1960	0.1930

Table 4.5: Fit Test for Low-ability IHS exam

Statistic	**Value**	**Std.Dev.**	p_{χ^2}	$p_{*\mathcal{X}}$	$p'_{*\mathcal{X}}$	p_{*M}
MNSQ_{out}	1.0039		0.3520			
$D^2_{out;\mathcal{X}}$	1.0047	0.0269				
$D^2_{out;\mathcal{Y}}$	1.0012	0.0347		0.4550	0.4590	0.4660
$D^2_{out;\mathcal{P}}$	1.0007	0.0220		0.4490	0.3820	0.3940
MNSQ_{in}	0.9959		0.4532			
$D^2_{in;\mathcal{X}}$	0.9956	0.0174				
$D^2_{in;\mathcal{Y}}$	0.9996	0.0184		0.4040	0.4260	0.4320
$D^2_{in;\mathcal{P}}$	1.0001	0.0066		0.3840	0.2490	0.2650

Table 4.6: Fit Test for High-ability IHS exam

Comparison of the three distributions (Figures A.1, A.2 and A.3) shows that while the original test is quite symmetric the two smaller tests are negatively and positively skewed, respectively.

For the total test both types of tests indicate model fit. MC tests, however, have larger p-values than the MNSQ outfit test.

On the infit side, in Table 4.4 MNSQ again shows the usual strong fit due to the effective blindness of infit. The MC indices, however, resulted in smaller p-values.

An interesting result can be derived from the fit tests of the two smaller exams. The two tests display excellent fit from every investigated point of view. Both traditional MNSQ p-values are very high (this is meaningful only for outfit as infit is almost incapable of showing anything but excellent model fit). The MC p-values are also very high. This is in accordance with the the result found in the global analysis, which indicated quite a strong relationship between MNSQ_{outfit} and $MC_{\mathcal{P}M}$. Even though, this relationship is not at all understood at this point the present analysis provides yet another indication of it.

The explanation of the very good fit of these data may be attributed to the fact that by splitting the test, the structure of the response matrix changed dramatically.

Note, however, that the distributional dependence study is concluded by noting that there is no observable difference between the fit analyses of the three data sets with different distributional properties.

To really perform a distribution dependence study one should consider the following global analysis. First, a large sample should be drawn from the collection of

response matrices with the same marginal sums (same marginal sums will of course imply distributional equality). This is a quite well understood process described in the literature (Snijders, 1991; Rao et al., 1996; Roberts, 2000; Ponoczny, 2001). The main computational tool is the Markov chain Monte Carlo method. For this sample a Monte Carlo fit test is performed. Finally, the entire process is repeated for *different score distributions.*

As explained, there is no real reason to go into this very involved algorithm. The global fit analysis of the previous section, however, has a much more interesting research question which, in a way, can be understood as a good and meaningful substitute for the distributional relationship study.

4.5 Stability Study

The goal of this study is to find an empirical threshold for K, a number K_t such that for all $K > K_t$ the fit analysis gives essentially the same result. This threshold is desired to be as small as possible, since the computational time increases with K.

Due to the nature of Monte Carlo method p-values provided by the new fit tests might show a certain degree of fluctuation over trials. Even under exactly identical circumstances, the corresponding D^2 and p-values will be different. However, only small differences are expected if the number of simulated matrices (K) is large enough.

For four different exam data the MC p-values were calculated with $K = 1000$ and with $K = 5000$, as well. Both cases were repeated five times. The stability study is performed by analyzing the obtained p-values.

K=1000					K=5000				
Repl\Exam	82×48	40×48	40×24	82×24	Repl\Exam	82×48	40×48	40×24	82×24
$p_{\mathcal{P}M1}$	0.3100	0.1810	0.3260	0.2260	$p_{\mathcal{P}M1}$	0.3200	0.1818	0.2988	0.2338
$p_{\mathcal{P}M2}$	0.3240	0.1740	0.2900	0.2260	$p_{\mathcal{P}M2}$	0.3188	0.1840	0.3082	0.2240
$p_{\mathcal{P}M3}$	0.2880	0.1800	0.3090	0.2080	$p_{\mathcal{P}M3}$	0.3218	0.1804	0.2976	0.2320
$p_{\mathcal{P}M4}$	0.3270	0.1620	0.3100	0.2410	$p_{\mathcal{P}M4}$	0.3204	0.1742	0.2960	0.2342
$p_{\mathcal{P}M5}$	0.3340	0.1770	0.2900	0.2390	$p_{\mathcal{P}M5}$	0.3298	0.1756	0.3024	0.2252
Mean	0.3166	0.1748	0.3050	0.2280	Mean	0.3222	0.1792	0.3006	0.2298
Std.Dev.	0.0182	0.0077	0.0153	0.0132	Std.Dev.	0.0044	0.0042	0.0049	0.0049
Range	0.0460	0.0190	0.0360	0.0330	Range	0.0110	0.0098	0.0122	0.0102
Minimum	0.2880	0.1620	0.2900	0.2080	Minimum	0.3188	0.1742	0.2960	0.2240
Maximum	0.3340	0.1810	0.3260	0.2410	Maximum	0.3298	0.1840	0.3082	0.2342

Table 4.7: Stability of $MC_{\mathcal{P}M}$: five replicated p-values of $MC_{\mathcal{P}M}$ for four different Math exams.

The following tables (Tables 4.7 through 4.12) display the respective Monte Carlo p-values resulting from five independent replications. The tables are in pairs. The first one always stands for the case when the number of simulated matrices is $K = 1000$, while the second contains the results from $K = 5000$ simulations.

These tables also contain the means of the Monte Carlo p-values along with their standard deviations, ranges, minima, and maxima.

The stability analysis shows very satisfactory results. Even though the range of $MC_{\mathcal{P}M}$ p-values reaches 0.0460 (for sample size $K = 1000$; test 82×48) it drops below 0.012 for all tests when the sample size is raised to $K = 5000$. The means show

K=1000					K=5000				
Repl\Exam	82×48	40×48	40×24	82×24	Repl\Exam	82×48	40×48	40×24	82×24
$p_{\mathcal{Y}M1}$	0.3710	0.2640	0.3900	0.3220	$p_{\mathcal{Y}M1}$	0.3856	0.2660	0.3652	0.3366
$p_{\mathcal{Y}M2}$	0.3820	0.2620	0.3600	0.3220	$p_{\mathcal{Y}M2}$	0.3804	0.2772	0.3688	0.3142
$p_{\mathcal{Y}M3}$	0.3650	0.2700	0.3820	0.3240	$p_{\mathcal{Y}M3}$	0.3786	0.2732	0.3710	0.3244
$p_{\mathcal{Y}M4}$	0.3760	0.2590	0.3530	0.3230	$p_{\mathcal{Y}M4}$	0.3774	0.2660	0.3648	0.3252
$p_{\mathcal{Y}M5}$	0.3810	0.2650	0.3650	0.3360	$p_{\mathcal{Y}M5}$	0.4038	0.2842	0.3730	0.3184
Mean	0.3750	0.2640	0.3700	0.3254	Mean	0.3852	0.2733	0.3686	0.3238
Std.Dev.	0.0071	0.0041	0.0155	0.0060	Std.Dev.	0.0109	0.0078	0.0036	0.0085
Range	0.0170	0.0110	0.0370	0.0140	Range	0.0264	0.0182	0.0082	0.0224
Minimum	0.3650	0.2590	0.3530	0.3220	Minimum	0.3774	0.2660	0.3648	0.3142
Maximum	0.3820	0.2700	0.3900	0.3360	Maximum	0.4038	0.2842	0.3730	0.3366

Table 4.8: Stability of $MC_{\mathcal{Y}M}$: five replicated p-values of $MC_{\mathcal{Y}M}$ for four different Math exams.

very convincing stability, as well. The largest jump is $0.3222 - 0.3166 = 0.006$, that is 0.6% (in the case of the largest, 82×48, response matrix).

For $MC_{\mathcal{Y}M}$, the fluctuation is even smaller in terms of the range for $K = 1000$, however, for increased number of simulated matrices $MC_{\mathcal{P}M}$ shows better stability. The largest range for $MC_{\mathcal{Y}M}$ with $K = 1000$ is 0.0370 (for 40×24 matrix) which falls significantly back to 0.0082 for $K = 5000$ matrices.

Ranges of $MC_{\mathcal{P}\mathcal{X}}$ are also small, but $p'_{P\mathcal{X}}$ has relatively large ranges for $K = 1000$ such as 0.0730 for 82×48 matrix and 0.0550 for the 40×24 one. The increased number of simulated matrices provides acceptable ranges for $MC_{\mathcal{P}\mathcal{X}}$, though.

K=1000					K=5000				
Repl\Exam	82×48	40×48	40×24	82×24	Repl\Exam	82×48	40×48	40×24	82×24
$p_{\mathcal{PX}1}$	0.3040	0.2270	0.3010	0.2340	$p_{\mathcal{PX}1}$	0.2970	0.2194	0.2874	0.2536
$p_{\mathcal{PX}2}$	0.3280	0.2040	0.3020	0.2320	$p_{\mathcal{PX}2}$	0.2976	0.2312	0.2924	0.2366
$p_{\mathcal{PX}3}$	0.2950	0.2360	0.3020	0.2470	$p_{\mathcal{PX}3}$	0.3120	0.2232	0.2940	0.2584
$p_{\mathcal{PX}4}$	0.3040	0.2250	0.2940	0.2440	$p_{\mathcal{PX}4}$	0.3054	0.2244	0.2954	0.2512
$p_{\mathcal{PX}5}$	0.3090	0.2360	0.3030	0.2740	$p_{\mathcal{PX}5}$	0.3072	0.2300	0.2898	0.2540
Mean	0.3080	0.2256	0.3004	0.2462	Mean	0.3038	0.2256	0.2918	0.2508
Std.Dev.	0.0123	0.0131	0.0036	0.0168	Std.Dev.	0.0064	0.0049	0.0032	0.0083
Range	0.0330	0.0320	0.0090	0.0420	Range	0.0150	0.0118	0.0080	0.0218
Minimum	0.2950	0.2040	0.2940	0.2320	Minimum	0.2970	0.2194	0.2874	0.2366
Maximum	0.3280	0.2360	0.3030	0.2740	Maximum	0.3120	0.2312	0.2954	0.2584

Table 4.9: Stability of $MC_{\mathcal{PX}}$: five replicated p-values of $MC_{\mathcal{PX}}$ for four different Math exams.

In summary, the question of stability shows remarkable results overall. Except for the p- and p'-values of $MC_{\mathcal{YX}}$ (which are really the same cases) the range of p-values with sample size $K = 5000$ is rarely over 2%! It is worthwhile to note once more, that a sample of size $K = 5000$ is an extremely small sample. The number of response matrices with size 82×48 is $2^{3936} \approx 7 \cdot 10^{1185}$. It is almost unbelievable that a sample from this set of size 5000 (or in many cases even 1000) produces stable result. This is due to the way this humongous set is sampled. Those matrices which are likely to be sampled are the ones with high probability (with respect to the measure $\mathfrak{p}$). Most matrices in this set come with negligible probability, even when compared to the size of the set M_{resp}.

K=1000					K=5000				
Repl\Exam	82×48	40×48	40×24	82×24	Repl\Exam	82×48	40×48	40×24	82×24
$p'_{\mathcal{P}\mathcal{X}1}$	0.3230	0.1810	0.3420	0.2110	$p'_{\mathcal{P}\mathcal{X}1}$	0.3174	0.1796	0.3000	0.2358
$p'_{\mathcal{P}\mathcal{X}2}$	0.3300	0.1740	0.2870	0.2140	$p'_{\mathcal{P}\mathcal{X}2}$	0.3168	0.1838	0.3098	0.2116
$p'_{\mathcal{P}\mathcal{X}3}$	0.2650	0.1810	0.3080	0.2130	$p'_{\mathcal{P}\mathcal{X}3}$	0.3268	0.1810	0.2976	0.2374
$p'_{\mathcal{P}\mathcal{X}4}$	0.3280	0.1620	0.3240	0.2240	$p'_{\mathcal{P}\mathcal{X}4}$	0.3178	0.1792	0.2982	0.2348
$p'_{\mathcal{P}\mathcal{X}5}$	0.3380	0.1810	0.2990	0.2540	$p'_{\mathcal{P}\mathcal{X}5}$	0.3162	0.1754	0.2938	0.2314
Mean	0.3168	0.1758	0.3120	0.2232	Mean	0.3190	0.1798	0.2999	0.2302
Std.Dev.	0.0295	0.0083	0.0215	0.0179	Std.Dev.	0.0044	0.0030	0.0060	0.0106
Range	0.0730	0.0190	0.0550	0.0430	Range	0.0106	0.0084	0.0160	0.0258
Minimum	0.2650	0.1620	0.2870	0.2110	Minimum	0.3162	0.1754	0.2938	0.2116
Maximum	0.3380	0.1810	0.3420	0.2540	Maximum	0.3268	0.1838	0.3098	0.2374

Table 4.10: Stability of $MC_{\mathcal{P}\mathcal{X}}$: five replicated p'-values of $MC_{\mathcal{P}\mathcal{X}}$ for four different Math exams.

It is easy to understand why $MC_{\mathcal{Y}\mathcal{X}}$ produces the least stable results. For, one only has to observe that the sampled response matrices are used the most intensively in this test. The p-value is calculated from two distributions, one which is the set of distances *among* simulated matrices which depends heavily on this simulated set. The other distribution depends on this set as well, since it is the set of distances between $\mathcal{X}$ and the simulated set. Even in this case the largest range is only 3.14%. Since the p-values themselves are all larger than 27%, this deviation is relatively small, and in no way can affect the conclusion of the hypothesis tests.

On the other hand the $MC_{\mathcal{P}M}$ and $MC_{\mathcal{Y}M}$ are the most desirable from the perspective of stability. For, both of them use the simulated set only once. Moreover,

K=1000					K=5000				
Repl\Exam	82×48	40×48	40×24	82×24	Repl\Exam	82×48	40×48	40×24	82×24
$p_{\mathcal{Y}\mathcal{X}1}$	0.2960	0.2250	0.2910	0.2410	$p_{\mathcal{Y}\mathcal{X}1}$	0.2986	0.2192	0.2802	0.2536
$p_{\mathcal{Y}\mathcal{X}2}$	0.3220	0.2080	0.2900	0.2320	$p_{\mathcal{Y}\mathcal{X}2}$	0.2996	0.2306	0.2836	0.2402
$p_{\mathcal{Y}\mathcal{X}3}$	0.2990	0.2360	0.3160	0.2520	$p_{\mathcal{Y}\mathcal{X}3}$	0.3106	0.2226	0.2920	0.2612
$p_{\mathcal{Y}\mathcal{X}4}$	0.2980	0.2250	0.2800	0.2560	$p_{\mathcal{Y}\mathcal{X}4}$	0.3060	0.2178	0.2922	0.2494
$p_{\mathcal{Y}\mathcal{X}5}$	0.3000	0.2300	0.3030	0.2690	$p_{\mathcal{Y}\mathcal{X}5}$	0.3196	0.2388	0.2894	0.2452
Mean	0.3030	0.2248	0.2960	0.2500	Mean	0.3069	0.2258	0.2875	0.2499
Std.Dev.	0.0107	0.0104	0.0138	0.0142	Std.Dev.	0.0086	0.0088	0.0053	0.0080
Range	0.0260	0.0280	0.0360	0.0370	Range	0.0210	0.0210	0.0120	0.0210
Minimum	0.2960	0.2080	0.2800	0.2320	Minimum	0.2986	0.2178	0.2802	0.2402
Maximum	0.3220	0.2360	0.3160	0.2690	Maximum	0.3196	0.2388	0.2922	0.2612

Table 4.11: Stability of $MC_{\mathcal{Y}\mathcal{X}}$: five replicated p-values of $MC_{\mathcal{Y}\mathcal{X}}$ for four different Math exams.

$MC_{\mathcal{P}M}$ appears to be the most natural fit index from the geometrical point of view (described in Chapter 3).

The only case when the small inaccuracy of MC tests can have effect on the final conclusion of the hypothesis test is when the p-value is around the prescribed significance level α.

K=1000					K=5000				
Repl\Exam	82×48	40×48	40×24	82×24	Repl\Exam	82×48	40×48	40×24	82×24
$p'_{\mathcal{Y}\mathcal{X}1}$	0.3750	0.2660	0.3960	0.3120	$p'_{\mathcal{Y}\mathcal{X}1}$	0.3842	0.2644	0.3654	0.3386
$p'_{\mathcal{Y}\mathcal{X}2}$	0.3840	0.2610	0.3590	0.3190	$p'_{\mathcal{Y}\mathcal{X}2}$	0.3792	0.2760	0.3696	0.3072
$p'_{\mathcal{Y}\mathcal{X}3}$	0.3530	0.2700	0.3820	0.3260	$p'_{\mathcal{Y}\mathcal{X}3}$	0.3816	0.2748	0.3708	0.3292
$p'_{\mathcal{Y}\mathcal{X}4}$	0.3770	0.2590	0.3590	0.3100	$p'_{\mathcal{Y}\mathcal{X}4}$	0.3752	0.2698	0.3658	0.3256
$p'_{\mathcal{Y}\mathcal{X}5}$	0.3850	0.2730	0.3730	0.3480	$p'_{\mathcal{Y}\mathcal{X}5}$	0.3956	0.2834	0.3694	0.3218
Mean	0.3748	0.2658	0.3738	0.3230	Mean	0.3832	0.2737	0.3682	0.3245
Std.Dev.	0.0129	0.0059	0.0158	0.0153	Std.Dev.	0.0077	0.0071	0.0024	0.0115
Range	0.0320	0.0140	0.0370	0.0380	Range	0.0204	0.0190	0.0054	0.0314
Minimum	0.3530	0.2590	0.3590	0.3100	Minimum	0.3752	0.2644	0.3654	0.3072
Maximum	0.3850	0.2730	0.3960	0.3480	Maximum	0.3956	0.2834	0.3708	0.3386

Table 4.12: Stability of $MC_{\mathcal{Y}\mathcal{X}}$: five replicated p'-values of $MC_{\mathcal{Y}\mathcal{X}}$ for four different Math exams.

CHAPTER 5

CONCLUSIONS

The primary goal of this work is to introduce a new method for fit analysis that is based on a Monte Carlo method. This new technique has been formulated and developed for the Rasch model and it has manifested itself in four different fit tests. The new technique combines the element of computer simulation, statistical hypothesis testing and Rasch model fitting. The innovative element in the proposed fit analysis is the creation of the sampling or model distributions "on the fly" which makes these tests non-parametric in nature. Fit indices that were proposed and used extensively in practice for the Rasch model (most notably MNSQ_{infit}, MNSQ_{outfit}) have been shown to have serious deficiencies due to assumed distributional properties of their fit statistics. In cases when their assumptions fail to hold, the appropriateness of fit analysis is in danger.

The non-parametric family of Monte Carlo fit tests introduced in this dissertation appears to be a good candidate for replacing the faulty MNSQ fit tests. The family contains several members among which no preference was given as of yet. Further study is planned to identify (if any) significant differences among the newly introduced tests. All of them acquired a solid theoretical foundation here, therefore they can be used with confidence.

A comparison study targeted to contrast MNSQ and MC indices confirms the results of earlier studies by showing significant departure from normality for MNSQ. Recall, that in all instances when the normality assumption holds, MNSQ fit tests are correct and their p-value should be taken seriously. Since the newly proposed indices involve no additional assumption, they deliver reliable fit information every time they are used. When the MNSQ normality assumption is appropriate the two indices should convey the same message about model fit. On the other hand, their significant difference can only be attributed to the in-appropriateness of the normality assumption. Since significant difference was found in this dissertation, another manifestation of the non-normality phenomenon should be recorded. None of the tests showed significant misfit for the response matrices under consideration in this study (and there were many thousands of them!), contradicting the traditional MNSQ outfit tests on several occasions.

Although the Monte Carlo p-values are always approximate, they show remarkable stability with respect to the number of simulated matrices. It is shown that for moderately sized tests a sample of size 5000 gives satisfactory result. This means that from the set of all response matrices an extremely small sample of K=5000 elements is satisfactory to model the entire probability distribution. This is in itself a remarkable finding which is due mainly to the careful choice of the simulation method. This fact also makes the introduction of more developed sampling methods unnecessary. Also, one has to keep in mind that the p-values - the final results of the analysis of fit - are calculated as relative frequencies. They are calculated as the ratio of the number

of sampled matrices with worse fit characteristic than the original response matrix and the number of simulated matrices. Hence, a p-value with $K = 1000$ simulated matrices can not be more accurate than $0.001 = \frac{1}{1000}$. For $K = 5000$ the maximal accuracy of the p-value increases to $0.0002 = \frac{1}{5000}$. As a consequence, no reasonable precision can be achieved with a small number of simulated matrices.

It is advised to always perform the stability study. It is highly unlikely that a one-size fits all answer can be given concerning the number of simulated matrices needed to obtain a reliable result. Some replications of the calculation of the Monte Carlo p-values, however, could easily lead to reasonably well founded decision on fit. In this study 5 replications were made for both sample sizes $K = 1000$ and $K = 5000$ and that was seen to be sufficient under the circumstances.

The results of the global fit analysis of the Rasch model show that when fit analysis is performed correctly the Rasch model performs very well. Indeed, small to medium sized examples show that the global fit index of the Rasch model is very small for $\alpha = 0.1$ and cannot be seen to differ from 0 at $\alpha = 0.05$. This means that the maximum number of misfitting matrices was 1 when the number of matrices sampled was 1000. Again, this sample is very small compared to the number of all response matrices, but the sampling was performed so as to yield intuitively misfitting matrices (that is, a very large number of possible aberrant responses). This choice was successful, as shown by the overwhelming number of matrices displaying strong MNSQ misfit.

In the light of this observation, one may arrive at the very pleasing conclusion, namely *the data almost always fit the Rasch model.* This could settle a long-standing

and unpleasant paradox of the Rasch model which is kept alive by the fact that even though the Rasch model is theoretically a most reasonable model it shows misfit for a considerably large number of tests. This dissertation shows that this was due to the violation of the assumptions of *model fit* and not to the model itself. In the light of these findings the Rasch model appears as an even more appealing model for test development.

Note, however, all results presented here are subject to the specific sampling of the set of response matrices. Thus, a straightforward extension of the present work could be a study of different sampling schemes represented by the choice of the generating matrix $\mathcal{G}$ using different IRT models. Also, it is an interesting question if the small matrices always show acceptable fit (as indicated by this study) or this was just a deficiency of the simulation scheme. The global fit analysis study could, in the long run, make good use of the advanced sampling methods (e.g., the Monte Carlo Markov Chain method). The one million simulations needed to create just one chart in the global analysis makes this kind of research very time consuming. Each figure was obtained by an approximately two day run on a well equipped up-to-date computer.

Another interesting observation is that there is a quite well displayed functional relationship between the p-values of MNSQ and $MC_{\mathcal{P}M}$. The question one might ask is if it is possible to identify this function in general. If so, then one could by-pass the calculation of the Monte Carlo fit index as it would be derivable from the MNSQ p-value (which is much easier to calculate).

Let us also remark on the running time of the tests. Since the Monte Carlo tests substitute the analytic solution of the fit problem with simulation it is computer

intensive. The code that was used in this study (Antal & Antal, 2003b) was written in a way to consume as little memory as possible. The running time for the largest test of size 82×48 with sample size 5000 took approximately 10 minutes on a machine with a 2.4 GHz Xenon processor. Knowing that a large group of item response theorists left the Rasch model because it unable to display significant model fit frequently (and misleadingly), the increased computational time should be worth clarifying the fit of a particular model. As high computing power becomes more and more commonplace there is nothing to obstruct the way of fit analysis proposed here.

There are two major directions that are expected to be followed in the future. One is to investigate the possibilities of adapting the proposed method to other IRT models, as well as to different kinds of test indices.

An IRT model is fully represented by its response probability matrix $\mathcal{P}$ with elements being the probabilities of the actual performance levels given all model parameters. Any IRT model defines the set of response matrices, that is the set of matrices with size $N \times L$ (N and L are as before) and with entries according to the model. As indicated earlier, the simulation method should be changed in accordance with the model. To sample the set of response matrices the relative frequency of a matrix has to be equal to its probability calculated from $\mathcal{P}$. It is also attempting to generalize the new technique to other fit statistics given by Equations 3.34 and 3.35. Background information on this research problem can be found in the last section of Chapter 3. The first addition to the current work in this direction could be the development of item and person MC tests.

Appendix A

FREQUENCY DISTRIBUTIONS OF IHS EXAMS

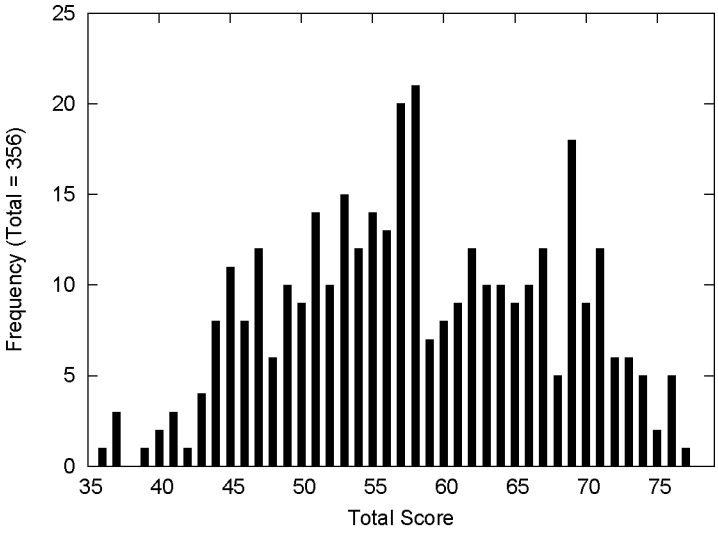

Figure A.1: Total Score Distribution for IHS Exam

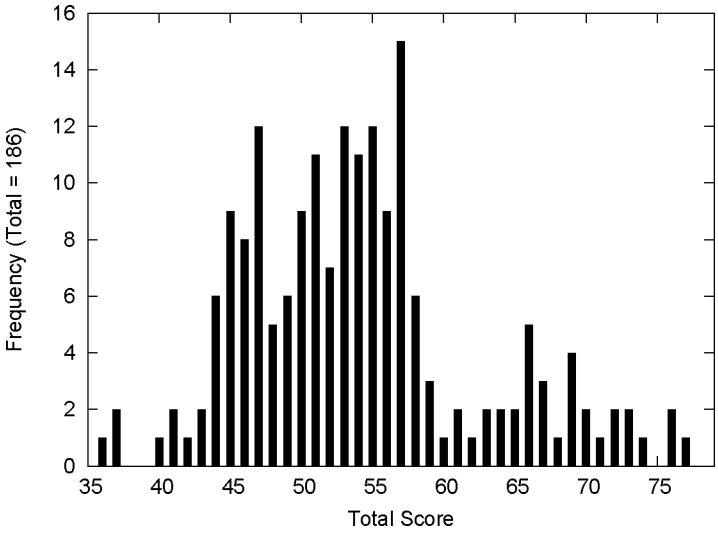

Figure A.2: Low-ability Portion of IHS exam

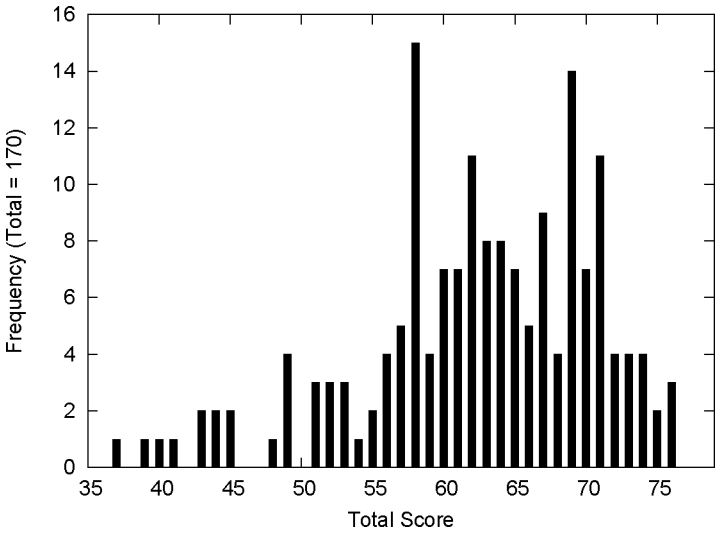

Figure A.3: High-ability Portion of IHS exam

Appendix B

FREQUENCY DISTRIBUTIONS OF GLOBAL FIT ANALYSIS

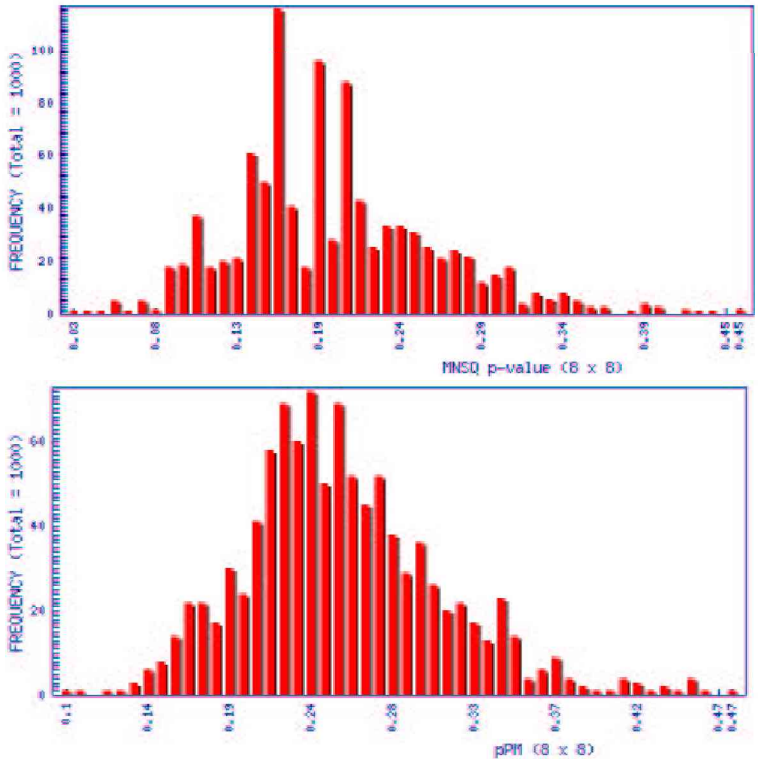

Figure B.1: Frequency distribution of the MNSQ p-value and p_{PM} (8×8).

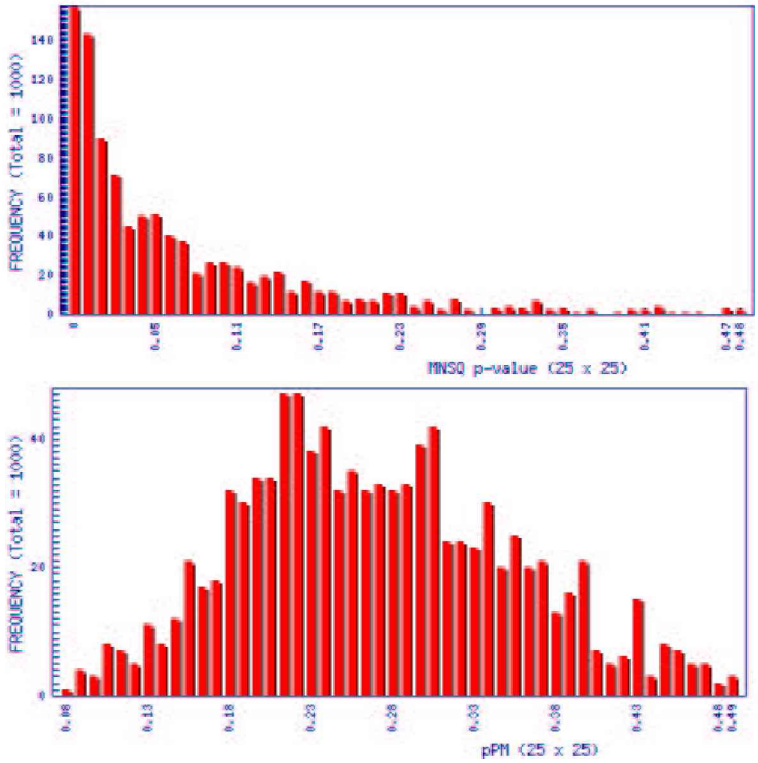

Figure B.2: Frequency distribution of the MNSQ p-value and p_{PM} (25×25).

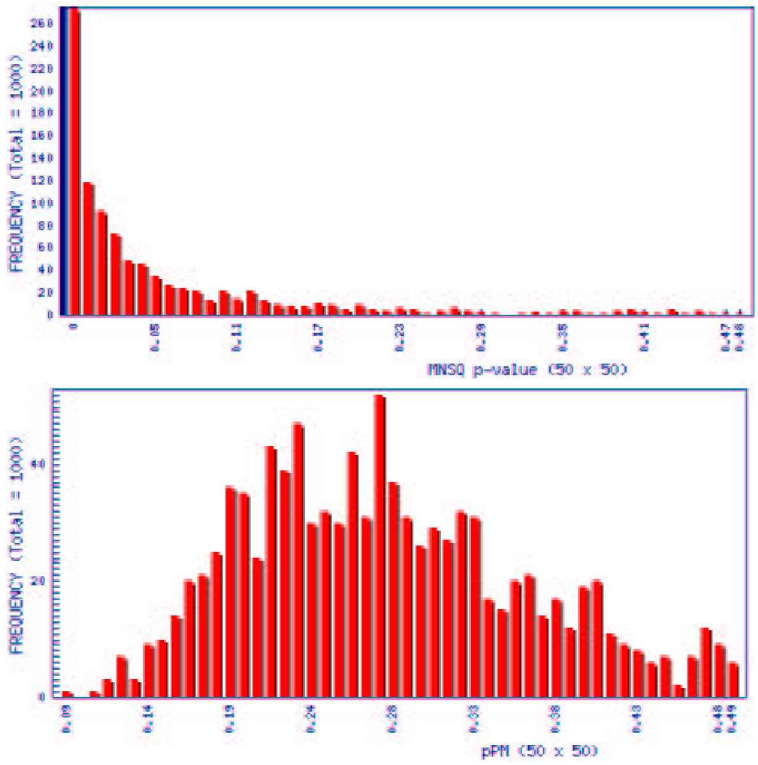

Figure B.3: Frequency distribution of the MNSQ p-value and p_{PM} (50×50).

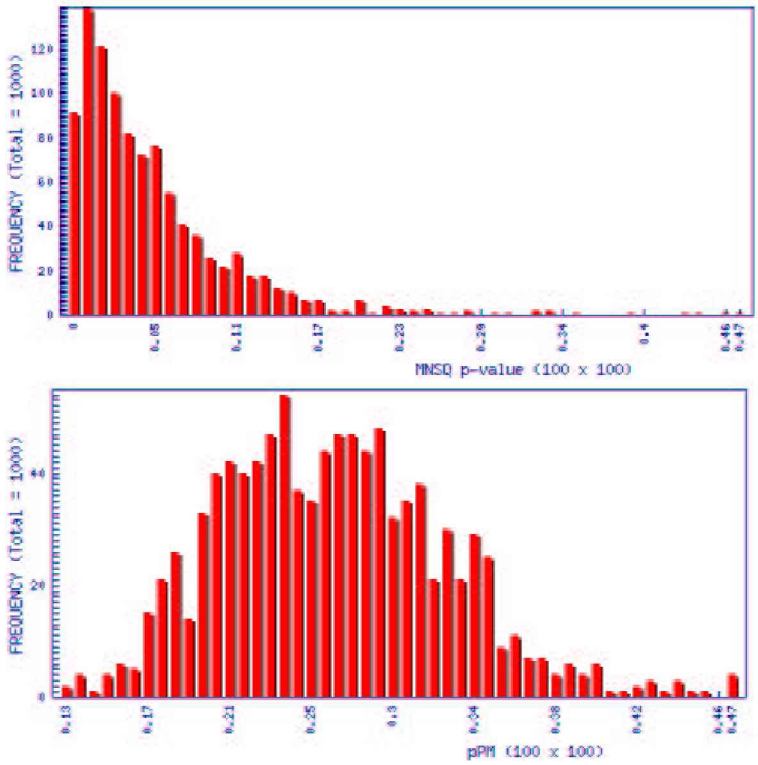

Figure B.4: Frequency distribution of the MNSQ p-value and p_{PM} (100×100).

Appendix C

SIMULATION OF RESPONSE MATRICES (PERL COMPUTER CODE)

```
# random(\@RASCH) returns random score matrix according
# to the Rasch probability matrix.
sub random {
        my $ref = $_[0];
        my @P   = @$ref;
        my ($i,@score, $P_rand, $n);
        for(my $j=0; $j<@P;$j++){
                $P_rand  = rand(1);
                $score[$j]=0 if $P_rand >  $P[$j];
                $score[$j]=1 if $P_rand <= $P[$j];
        }
        return (@score);
}
```

BIBLIOGRAPHY

Allen, M. & Yen, W. (1979). *Introduction to Measurement Theory.* Monterey, CA: Brooks/Cole.

Antal, T & Antal, J. (2003) Kardinál 0.018, Rasch IRT software.

Baker, F. (1992). *Item Response Theory: Parameter estimation methods.* New York, NY: Marcel Dekker, Inc.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & Novick (Eds.)Statistical theories of mental test scores (pp. 397-472). Reading, MA: Addison-Wesley.

Dorans, N. J., & Schmitt, A. P. (1993). Structured response and differential item functioning: A programmatic perspective. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135-165). Hillside, NJ: Erlbaum.

Drasgow, F., Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10,* 59-67.

Drasgow, F., Levine, M. V. (1985). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171-191.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 59-79.

Drasgow, F., Levine, M. V. & Williams, E. A. (1985). Appropriateness measures with polythomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

D'Costa, A. G. (1993). Extending the Sato Caution Index to define the within and beyond ability caution indexes. Paper presented at the Annual Meeting of the National Council for measurement in Education. Atlanta, GA.

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society. 51,* 599-635.

Edgeworth, F. Y. (1892). Correlated averages. *Philosophical Magazine,* 5th Series, 34, 190-204.

Embretson, S. E. & Hershberger, S. L. (1999). *The new rules of measurement. What every psychologist and educator should know.* Lawrence Erlbaum Associate, NJ.

Emons, W.; Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 person-fit statistic. *Applied Psychological Measurement, 26,* 1, 88-108.

Fisher, G. H. (1981). On the existence and uniqueness of maximum likelihood estimates in the Rasch model. *Psychometrika, 46,* 59-77.

Ferrara, S., Huynh, H. & Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large-scale performance assessment. *Applied Measurement in Education 10,* 2, 123-144.

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

Harnisch D. L. & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133-146.

Holland, P. W. & Wainer, H. (Eds.), (1993). Differential item functioning. Hillsdale, NJ: Erlbaum.

Hopkins, K. D. (1998). *Educational and Psychological Measurement and Evaluation.* 8th Ed. Boston: Allyn and Bacon.

Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 97-110). New York, NY: Springer.

Kress, R. (1993). *Numerical Analysis.* New York, NY: Springer.

Lang, S. (1986). *Introduction to Linear Algebra.* New York, NY: Springer.

Levine, M. V. & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43,* 675-685.

Levine, M. V. & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35,* 42-56.

Levine, M. V. & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait and computerized adaptive testing* (pp. 109-131.) New York: Academic Press.

Li, Mao-neng F. & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21,* 3, 215-231.

Linacre, J. M. (1997). An All purpose person statistic? Rasch Measurement Transactions 11:3 p. 582-3.

Linacre, M. (1998). *BIGSTEPS User Guide.* Chicago: MESA Press.

McCamey, R. (2002). A primer on the one-parameter Rasch model. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, TX.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina and Whitney study. *Applied Psychological Measurement, 21,* 99-113.

Meijer, R. R. (1998). Consistency of test behaviour and individual difference in precision. *Journal of Occupational and Organizational Psychology, 71,* 147-160.

Meijer, R. R. & Sijtsma, K. (2002). Methodology review: Evaluating person fit. *Applied Psychological measurement. 25,* 2, 107-135.

Mislevy, R. J. (1994). Test theory reconvened. CSE Technical report 376. National Center for research on Evaluation, Standards, and Student Testing. UCLA CA: Los Angeles.

Molenaar, I. W. & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55,*75-106.

Myung, I. J. (in press). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology.*

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19,* 121-129.

Nering, M. L. (1997). The distribution of indexes of person-fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21,* 115-127.

Noonan, B. W., Boss. M. W. & Gessaroli, M. E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement, 16,* 4, 345-352.

Patel, J. K. & Reed, C. B. (1996). *The handbook of the normal distribution.* 2nd Edition, Revised and Expanded. New York, NY: Marcel Dekker, Inc.

Ponoczny, I. (2001). Nonparametric Goodness-of-fit tests for the Rasch model. *Psychometrika,* Vol 66, N 3, 437-460.

Rasch, G. (1960/80). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Danish Institute for Educational Research. Expanded edition (1980), with foreword and afterword by B. D. Wright. Chicago: University of Chicago Press.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19,* 213-229.

Reise, S. P. & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65,* 143-151.

Sato, T. (1980). The S-P Chart and the caution index. NEC Educational Information Bulletin, 80-1, C&C Systems Research Laboratories, Nippon Electronic Co., Ltd., Tokyo, Japan.

Schmitt, N.; Chan, D.; Sacco, J.M.; McFarland, L. A. & Jennings, D. (1999). Correlates of person fit and effect fit on test validity. *Applied Psychological Measurement, 23,* 41-53.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45,* 433-444.

Smith, R. M. (1995). Using item mean squares to evaluate fit to he Rasch model. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated response parameter. *Psychometrika, 66,* 3, 331-342.

Spearman, C. (1904). The proof of measurement of association between two things. *American Journal of Psychology,15,* 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18,* 161-169.

Stocking, M. L. (1989). Empirical estimation errors in item response theory as a function of test properties. Research Report RR-89-5. Princeton, NJ: ETS.

Swaminathan, H. (1983). Parameter estimation in item response theory models. In R. Hambleton (Ed.) *Applications of item response theory*(pp.24-44.) Vancouver, BC: Educational Research Institute of British Columbia.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95-110.

Tatsuoka, K. K. & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7,* 81-96.

Chen, Wen-Hung & Thissen, D. (1997). Local Dependence Indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22* 3.

Thorndike E. L., Bergman, E. O., Cobb, M. V. & Woodyard, E. (1926). *The measurement of intelligence.* New York: Teachers College Bureau of Publications.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16,* 433-451

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

Wainer, H., Dorans, n. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing.: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wall, L.; Christiansen, T. & Orwant, J. (2000). *Programming Perl.* Sebastopol, CA: O'Reilly.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,*

Wright, B. D. (1999). Fundamental measurement for Psychology. In: S. E. Embretson & Hershberger S. L. (Eds.) *The new rules of measurement.* pp. 65-105.

Wright, B. D. and Douglas, G. A. (1975a). *Best test design and self-tailored testing.* Research Memorandum No. 19, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D. and Douglas, G. A. (1975b). *Better procedures for sample-free item analysis.* Research Memorandum No. 20, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D. and Douglas, G. A. (1977a). Best procedures for sample-free item analysis. *Applied Psychological measurement, 1,* 281-294.

Wright, B. D. & Linacre, M. (1985). *Microscale manual* (ver. 2.0) Westport CT: Mediax Interactive Technologies.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis,* Rasch Measurement, Chicago: MESA Press.

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

Wright, B. D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement,* 29, 23-48.

Yen, W. M. (1994). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement, 30.*

Zickar, M. J. & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement.*Applied Psychological Measurement, 20,* 71-87.