THE OHIO STATE UNIVERSITY

QUANTITATIVE RESEARCH EVALUATION AND MEASUREMENT PROGRAM

ANTAL JUDIT

MONTE CARLO GOODNESS OF FIT TESTS FOR THE RASCH MODEL

Academic Advisor: Ayres D'Costa, Ph.D.

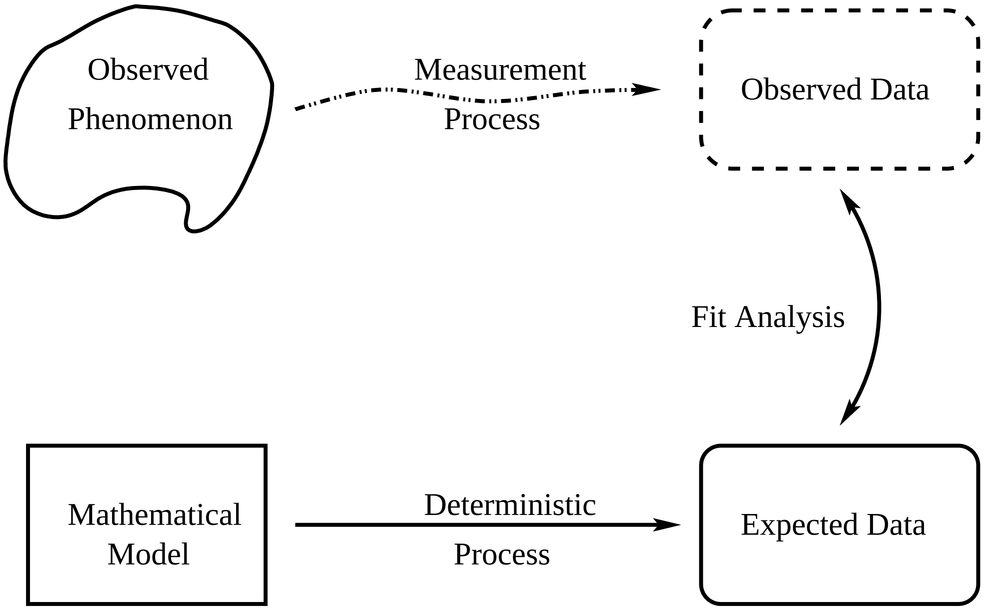
2003

1 Introduction

Test theory models are invented to understand the complicated interaction between test items and subjects taking the test. Mathematical models are always obtained by significantly simplifying real life situations, which also requires to keep a good balance between the complexity of the model and the agreement between the model and reality. The more complex a well crafted model is, the better it may describe the observed phenomenon. At the same time, however, complexity acts against our understanding. It is more difficult to capture the essence of a theory if it is made up of too many complex components.

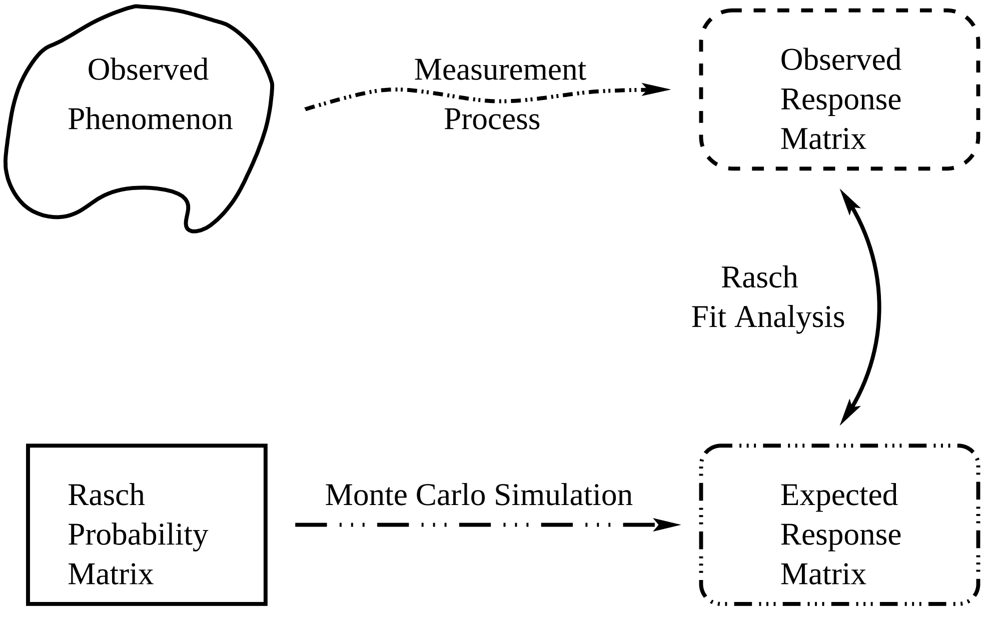
The ability of a model to explain the above mentioned interaction - which highly influences our confidence in the model - is measured by *goodness of fit indices.* That is why it is of extreme importance to produce highly trusted fit indices. One is reluctant to discard a beautiful theory and use a complicated one unless there is a substantial inconsistency between the simpler model and the data. The authors' impression is that models are frequently put on the shelf based on inappropriate and misleading fit indices. A highly simplified diagram representing the fit problem is shown in Figure 1.

Figure 1: Model Fit: Goodness of fit is determined by comparing observed data to expected data. The latter is determined by analytic methods.



Apart from the model selection aspect, there is another issue where serious model fit analysis comes into play. There are numerous studies targeting all sorts of testing behavior such as differential item functioning (DIF) (Zwick, 2002; Roussis, 1999; Parshall & Miller, 1995; Dorans & Smith, 1993; Holland & Wainer, 1993; Cohen et al, 1991; Swaminathan & Rogers, 1990) and local item dependence (LID) (Ferrara & Huynh, 1999; Reese, 1999; Ferrara, 1997; Wilson, 1988; Yen, 1983). The identification of DIF or LID is performed by introducing specific indices which produce significant departure from their expected values when DIF or LID is present. Many indices designed to indicate significant deviation from normal behavior rely on good model fit. Ponoczny (2001) presents great many examples of various indices that maybe targeted with the newly proposed

Figure 2: Monte Carlo model fit in the Rasch model: expected data is generated rather than derived by analytic methods.



technique. The frequently used IRT fit indices (e.g. mean square infit and outfit) all suffer from the same deficiency: their exact null distributions are unknown. There are practical assumptions made about these null distributions which are widely used, and widely criticized.

Since it is highly unlikely that exact analysis of the fit problem can ever be provided, for the time being one is left with simulation methods. The advantage of a simulation method is that it is relatively easy to implement and, to a certain degree, it is reliable. On the other hand, it is time consuming and computer resource intensive. Also, the resulting fit indices cannot be reproduced exactly. Nevertheless, it will be shown that this method is much more reliable than the previously used tests. Figure 2 displays the main components of the Monte Carlo fit test for the Rasch model proposed in this paper.

The structure of this paper is as follows. First the relevant terms used in the study will be introduced. This is 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 fit, laying down the foundation of the newly proposed technique. This general point of view is the one that lights the road to the "Monte-Carlo solution" of the fit problem. The next section will give a thorough account of the fundamental idea behind the new fit index. Then comes the core of the paper, introducing the new family of fit indices in its entirety. This part also contains a theoretical comparison of the general, the residual based, and the newly introduced fit analyses, as well as the full description of the simulation method used. In the concluding two sections the performance of the new tests will be investigated from different viewpoints using real test data.

2 Item Response Theory

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 matrix with entries 0 and 1, 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. The elements x_{ji} of $\mathcal{X}$ assume the value of 1 or 0. x_{ji} is 1 if subject j scored correctly on item i and 0 otherwise. The number of subjects and the number of items are denoted by N and L, respectively.

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

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

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 it will be argued, the single impediment 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 Rasch Model

The Rasch model (Rasch, 1960; Wright & Stone, 1979; Wright & Masters, 1982; Wright & Mok, 2000; Smith, 2001; Stone, 2001; Baker, 2001; Linacre & Wright, 2002;) 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 by ϑ_j, respectively.

Under the Rasch model 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:

$$\begin{aligned} P_{ji} \quad &:= \quad Prob(x_{ji} = 1 \mid \vartheta_j, \delta_i) = \frac{e^{\vartheta_j - \delta_i}}{1 + e^{\vartheta_j - \delta_i}} = \frac{1}{1 + e^{\delta_i - \vartheta_j}}, &(2)\\ Q_{ji} \quad &:= \quad Prob(x_{ji} = 0 \mid \vartheta_j, \delta_i) = 1 - P_{ji} = \frac{1}{1 + e^{\vartheta_j - \delta_i}}. &(3) \end{aligned}$$

The function P_{ji} (Equation 2) is called the *Rasch item characteristic function* (or *curve*); Figure 3 shows the Rasch ICF.

The *Rasch* or *response probability matrix* $\mathcal{P}$ is the matrix with elements P_{ji} (Equation 2).

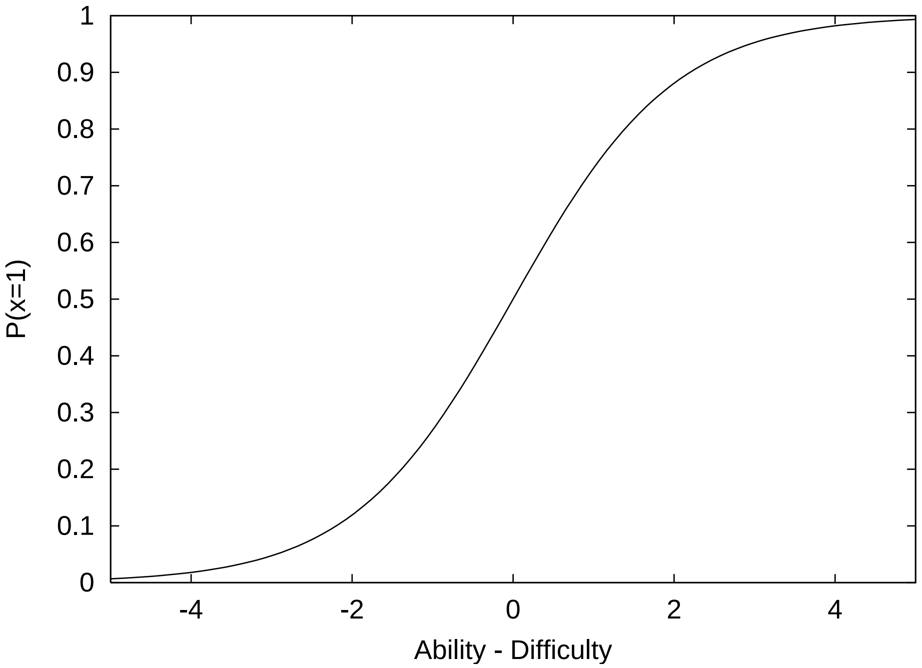
The conditional probability of one matrix element x_{ji} is then given by

$$P(x_{ji}) := \text{Prob}(x_{ji} \mid \delta_i, \vartheta_j) := \begin{cases} P_{ji} & \text{if} \quad x_{ji} = 1, \\ Q_{ji} = 1 - P_{ji} & \text{if} \quad x_{ji} = 0. \end{cases} \tag{4}$$

Also,

$$Q(x_{ji}) := 1 - P(x_{ji}). \tag{5}$$

Figure 3: 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 probabilities of each element of the item response matrix as in Equation 4. 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.

The second assumption already has been made use of 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*.

What is crucial for the theory is the conditional probability $L(\mathcal{X})$ 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 is the statistical independence of the matrix elements and it is expressed in the following equality

$$L(\mathcal{X}) := L(\mathcal{X}, \Delta, \Theta) := \mathrm{Prob}(\mathcal{X} \mid (\vartheta_j, \delta_i)_{1 \le i \le L, 1 \le j \le N}) = \prod_{(x_{ji}) = \mathcal{X}} P(x_{ji}). \tag{6}$$

To obtain the parameter estimates this total independence suffices. For the purposes of fit analysis, however, one has to introduce *local independence*. The latter is formulated by requiring that any set of the elements of $\mathcal{P}$ is be independent.

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 locally independent conditional probabilities $P(x_{ji})$ of the matrix element x_{ji}, given by (4);*
- *the conditional probability $L(\mathcal{X})$ of $\mathcal{X}$, given by (6).*

2.3 Joint Maximum Likelihood Estimation

Estimation procedures make parameter estimates available through an iteration procedure. The joint maximum likelihood estimation (JMLE) (Baker, 1992; de Leeuw & Verhelst, 1986; Fisher, 1981; Wright & Stone, 1979) procedure considers difficulties and abilities on the same footing. By definition, JMLE finds parameter estimates such that the likelihood function $L(\mathcal{X};\Delta,\Theta)$ corresponding to the response matrix $\mathcal{X}$ is maximal.

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

$$\mathcal{L} := \log(L(\mathcal{X})) = \sum_{(x_{ji})=\mathcal{X}} \log\left(\mathrm{P}(x_{ji})\right) = -\sum_{(x_{ji})=\mathcal{X}} \log\left(1+e^{(2x_{ji}-1)(\delta_i-\vartheta_j)}\right). \tag{7}$$

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

$$x_0 = (\delta_1^0, \delta_2^0, \ldots, \delta_L^0, \vartheta_1^0, \vartheta_2^0, \ldots, \vartheta_N^0) \tag{8}$$

for the iteration. The iteration scheme is then given by

$$x_{n+1} = x_n - \left(D^2\mathcal{L}(x_n)\right)^{-1} \cdot D\mathcal{L}(x_n), \tag{9}$$

where

$$x_n = (\delta_1^n, \delta_2^n, \ldots, \delta_L^n, \vartheta_1^n, \vartheta_2^n, \ldots, \vartheta_N^n) \tag{10}$$

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

The iteration stops when the Euclidean norm of the difference of two consecutive vectors $\|x_{n+1} - x_n\|$ is less than a prescribed number, the iteration margin.

For a well-conditioned response matrix this procedure has a unique solution (Fisher, 1981). In other words, the iteration scheme is converging to a unique finite 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. Note that uniqueness is achieved only after prescribing an extra condition (e.g. the mean of item difficulties is zero).

3 Model Fit

3.1 Mean Square Residual Tests

The commonly used fit statistics for the Rasch model are MNSQ outfit and infit (Wright & Stone, 1979; Meijer & Sijtsma, 2001).

The total MNSQ outfit is defined as

$$\text{MNSQ}_{out} = \frac{1}{NL}\sum_{i=1}^{L}\sum_{j=1}^{N}\frac{(x_{ji}-P_{ji})^2}{P_{ji}Q_{ji}}, \tag{11}$$

MNSQ infit is defined as

$$\text{MNSQ}_{in} = \frac{\sum\limits_{i=1}^{L}\sum\limits_{j=1}^{N}(x_{ji}-P_{ji})^2}{\sum\limits_{i=1}^{L}\sum\limits_{j=1}^{N}P_{ji}Q_{ji}}. \tag{12}$$

One defines the standard z score of MNSQ outfit as follows:

$$\text{MNSQ}_{zstd} := \sqrt{\frac{9NL}{2}}\left(\sqrt[3]{MNSQ}-1+\frac{2}{9NL}\right). \tag{13}$$

For infit one finds the standardized value by

$$\text{MNSQ}_{in,zstd} := \frac{3\left(\sqrt[3]{\text{MNSQ}_{in}}-1\right)}{q}-\frac{q}{3}, \tag{14}$$

where q is the standard deviation of infit given by

$$q = \frac{\sum\limits_{i=1}^{L}\sum\limits_{j=1}^{N}P_{ji}Q_{ji}(P_{ji}-Q_{ji})^2}{\sum\limits_{i=1}^{L}\sum\limits_{j=1}^{N}P_{ji}Q_{ji}} \tag{15}$$

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

According to the assumptions (Wright & Stone, 1979; Wright & Masters; 1982) both MNSQ_{zstd} follow a standard normal distribution $N(0,1)$.

A closer look at the MNSQ fit indices reveals that they are nothing else than squared distances between the Rasch probability matrix $\mathcal{P}$ and the response matrix $\mathcal{X}$ with respect to specific weights.

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

$$D_w(x,y) = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2 w_i}. \tag{16}$$

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

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

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

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

From Equation 11 it is clear that MNSQ_{out} is the distance squared

$$\mathrm{MNSQ}_{out} = D^2_{w_o}(\mathcal{P}, \mathcal{X}) \tag{17}$$

with weight

$$w_{o,ji} = \frac{1}{NL \cdot P_{ji} Q_{ji}}. \tag{18}$$

Similarly, for the MNSQ infit (Equation 12) one has

$$\mathrm{MNSQ}_{in} = D^2_{w_i}(\mathcal{P}, \mathcal{X}), \tag{19}$$

where this time the weight is

$$w_{i,ji} = \frac{1}{\sum\limits_{i=1}^{L} \sum\limits_{j=1}^{N} P_{ji} Q_{ji}}. \tag{20}$$

Note, that $w_{i,ji}$ is a constant weight vector.

3.2 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 one to pinpoint where the particular fit indices enter the picture and how their deficiency can be remedied.

In the framework of dichotomous IRT the number of 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 it 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. 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 $\mathcal{Y} = (y_{ji})_{j,i}$ be an element of $M_{resp}(N,L)$. The probability required for hypothesis testing concerning model fit is the conditional probability of $\mathcal{Y}$ given $\mathcal{P}$

$$\mathrm{Prob}(\mathcal{Y} \mid \mathcal{P}) := \prod_{ji} (y_{ji} P_{ji} + (1 - y_{ji}) Q_{ji}). \tag{21}$$

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

$$\mathfrak{p} := \mathfrak{p}(\mathcal{P}) : M_{resp}(N,L) \to [0,1] : \ \mathcal{Y} \mapsto \mathrm{Prob}(\mathcal{Y} \mid \mathcal{P}) \tag{22}$$

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 $\mathrm{Prob}(\mathcal{Y} \mid \mathcal{P})$ satisfying $\mathrm{Prob}(\mathcal{Y} \mid \mathcal{P}) < \mathrm{Prob}(\mathcal{X} \mid \mathcal{P}) = L(\mathcal{X})$ over $\mathcal{Y} \in M_{resp}(N,L)$, that is

$$p_o := \sum_{\mathcal{Y}:\ \mathrm{Prob}(\mathcal{Y} \mid \mathcal{P}) < \mathrm{Prob}(\mathcal{X} \mid \mathcal{P})} \mathrm{Prob}(\mathcal{Y} \mid \mathcal{P}). \tag{23}$$

By introducing

$$\mathcal{B}_L := \{\mathcal{Y} \ : \ \mathrm{Prob}(\mathcal{Y} \mid \mathcal{P}) < \mathrm{Prob}(\mathcal{X} \mid \mathcal{P})\} \tag{24}$$

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

$$p_o := \mathfrak{p}(\mathcal{B}_L) := \int\limits_{\mathcal{B}_L} 1 \ \mathrm{d}\mathfrak{p} \tag{25}$$

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

$$\mathcal{B}_{\text{MNSQ}} := \{\mathcal{Y} \ : \ D^2(\mathcal{Y},\mathcal{P}) \leq \mathrm{MNSQ}(\mathcal{X}) \leq 1\} \cup \{\mathcal{Y} \ : \ D^2(\mathcal{Y},\mathcal{P}) \geq \mathrm{MNSQ}(\mathcal{X}) \geq 1\} \tag{26}$$

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

$$p_o^{\text{MNSQ}} := \mathfrak{p}(\mathcal{B}_{\text{MNSQ}}) := \int\limits_{\mathcal{B}_{\text{MNSQ}}} 1 \ \mathrm{d}\mathfrak{p}. \tag{27}$$

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 $\mathfrak{p}$ by exact methods.

Note, that the probability p_o in Equation 25 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.

MNSQ 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 $[0.01, 0.1]$). 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).

4 A Family of Non-parametric Fit Indices

4.1 Statement of Problem

Unfortunately, both MNSQ outfit and infit use a strong assumption about their distributions. In cases when these assumptions fail to hold, the appropriateness of MNSQ tests is in danger. There are several studies targeting this issue (Li & Olejnik, 1997; Noonan, Boss & Gessaroli, 1992; Wright & Linacre, 1985; Smith, 1985). They all show evidence to that the distribution of MNSQ statistics depart from normality.

Li and Olejnik (1997) conducted a simulation study for $\mathrm{MNSQ}_{out,zstd}$ person fit with several test and misfit scenarios. For each scenario 50 replications were made resulting in a fit index distribution. Investigating the properties 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 $\mathrm{MNSQ}_{out,zstd}$ person fit shows a great departure from normality with large skewness and extremely large variable kurtosis.

Indications of these results were earlier addressed by Smith (1985) and Wright and Linacre (1985). Wright and Linacre suggested raising the cut-off value for rejecting person fit to 3.0 (from the standard 2.0) or beyond to compensate for non-normality.

4.2 Simulation of Rasch Response Matrices

In this section a simulation algorithm is described. It is used to generate Rasch response matrices used in the Monte Carlo fit tests. The initial data is the Rasch probability matrix $\mathcal{P}$ 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

$$y_{ji} = \begin{cases} 1 \text{ if } r \leq P_{ji}, \\ 0 \text{ otherwise.} \end{cases} \tag{28}$$

where P_{ji} is the Rasch probability as in Equation 2.

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

$$\mathcal{E}(\mathcal{Y}) = \mathcal{P} \quad \text{or} \quad \mathcal{E}(y_{ji}) = P_{ji} \ \forall (i,j). \tag{29}$$

The heart of the Monte Carlo 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 $\mathfrak{p}(\mathcal{Y}) = Prob(\mathcal{Y} \mid \mathcal{P})$.*

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.

The following theorem reveals the relationship of the several expected values appearing in the simulation. The proof is routine calculation and is left to the reader.

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

$$\mathcal{E}(y_{ji}) = P_{ji}. \tag{30}$$

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 (for both infit and outfit):

$$\mathcal{E}\left(D^2(\mathcal{X}, \mathcal{Y})\right) - MNSQ(\mathcal{X}) = 1, \tag{31}$$

$$\mathcal{E}\left(D^2(P, \mathcal{Y})\right) = 1, \tag{32}$$

$$\mathcal{E}\left(D^2(\mathcal{Y}, \mathcal{Y}')\right) = 2. \tag{33}$$

We close this section 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 non-parametric fit tests. Ponoczny chose the simulated matrices $\mathcal{Y}$ so 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). It is also argued by the same authors that the problem in general, due to the size of this set, is very hard to solve from practical 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 (Chen & Thissen, 1999), these are all approximative methods and use some parametric assumption about the underlying data - this is exactly what the Monte Carlo method wants to avoid.
- From the perspective of the probability measure (Equation 22) 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

  $$\frac{\mathfrak{p}(\mathcal{B} \cap C)}{\mathfrak{p}(C)} = \mathfrak{p}(\mathcal{B}), \tag{34}$$

  where $\mathcal{B}$ is any set introduced in conjunction with the fit analysis as in Equations 24 and 26. The rather strong independence condition in Equation 34 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.

4.3 Algorithm Defining the New Tests

This section presents the family of Monte Carlo (MC) fit indices. The inputs of this procedure are a response matrix $\mathcal{X} \in M_{resp}(N, L)$ and a natural number K, the number of simulated matrices. First, a long list of procedures is provided and then it will be shown how to combine them to obtain particular members of the family. One introduces the general notation D^2 for the weighted distances defined earlier.

(a) Run a joint maximum likelihood estimation (Wright and Stone (1979) p. 62; Baker (1992) p. 144) (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:

$$D^2_{k,out-\mathcal{X}} \quad := \quad D^2_{out}(\mathcal{X}, \mathcal{Y}_1^k) := \frac{1}{NL} \sum_{i=1}^{L} \sum_{j=1}^{N} \frac{\left(x_{ji} - y_{1,ji}^k\right)^2}{P_{ji} Q_{ji}}, \tag{35}$$

$$D^2_{k,in-\mathcal{X}} \quad := \quad D^2_{in}(\mathcal{X}, \mathcal{Y}_1^k) := \frac{\sum_{i=1}^{L} \sum_{j=1}^{N} \left(x_{ji} - y_{1,ji}^k\right)^2}{\sum_{i=1}^{L} \sum_{j=1}^{N} P_{ji} Q_{ji}}. \tag{36}$$

This provides us with 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:

$$D^2_{k,out-\mathcal{P}} \quad := \quad D^2_{out}(\mathcal{P}, \mathcal{Y}_2^k), \tag{37}$$

$$D^2_{k,in-\mathcal{P}} \quad := \quad D^2_{in}(\mathcal{P}, \mathcal{Y}_2^k). \tag{38}$$

This provides us with 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

$$D^2_{k,out-\mathcal{Y}} := D^2_{out}(\mathcal{Y}_2^{k'}, \mathcal{Y}_2^{k''}), \tag{39}$$
$$D^2_{k,in-\mathcal{Y}} := D^2_{in}(\mathcal{Y}_2^{k'}, \mathcal{Y}_2^{k''}), \tag{40}$$

where (k', k'') are randomly chosen pairs.
Consider the following hypothesis tests (for both outfit and infit; for simplicity we dropped the subscripts *out* and *in*):

(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 from different populations.

(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 from different populations.

Note: The algorithm makes use of the two simulated sets in tests (f) and (g) where they are needed two ensure that the two distributions involved 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 goodness of fit tests are developed using the previous list

$$MC_{\mathcal{P}\mathcal{X}} : (a) \to (b) \to (c) \to (d) \to (f), \tag{41}$$
$$MC_{\mathcal{Y}\mathcal{X}} : (a) \to (b) \to (c) \to (e) \to (g), \tag{42}$$
$$MC_{\mathcal{P}M} : (a) \to (b) \to (d) \to (h), \tag{43}$$
$$MC_{\mathcal{Y}M} : (a) \to (b) \to (e) \to (i). \tag{44}$$

To actually perform the hypothesis test one may chose the method of computing the one tail probability by counting the elements in one distribution that are higher (or lower, as the case maybe) than the mean of the other distribution. This results in two p-values (denoted by p and p') in the case of two distributions (as in (f) and (g)), and one in the case of a single distribution (as in (h) and (i)) (see Figures 4 and 5).

Figure 4: p-values from two distributions.

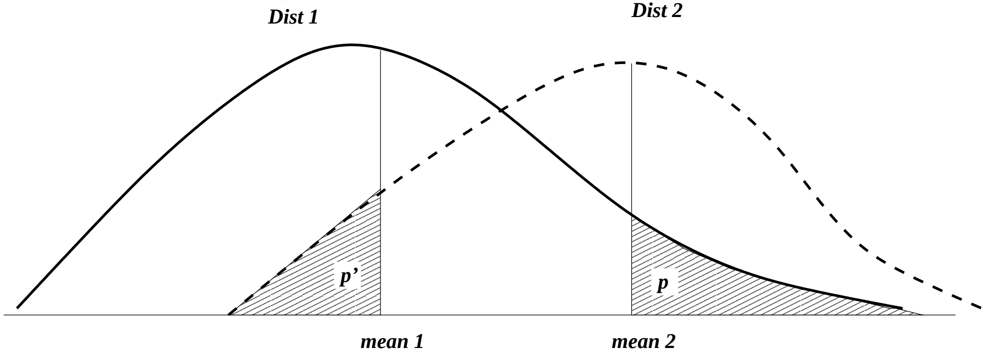
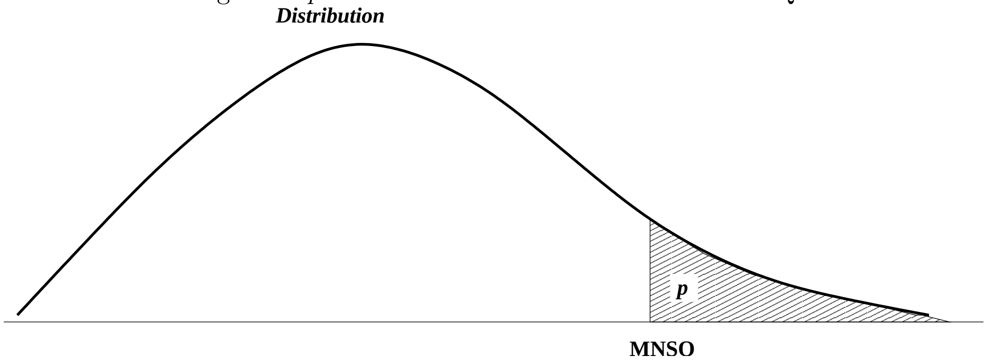


Figure 5: p-value from one distribution with MNSQ



4.4 Description and Specifics of the Tests

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 distribution is related to the model (which 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 $\mathcal{Y}$. This distribution, again, is related to the model via $\mathcal{P}$. 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 process itself. 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 later.

4.5 Fit Test MC$_{\mathcal{P}\mathcal{X}}$

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 this distance is close to 1. To make the comparison, the test MC$_{\mathcal{P}\mathcal{X}}$ "blows up" $\mathcal{P}$ and substitutes it with two sets of response matrices $\mathcal{Y}_1$ and $\mathcal{Y}_2$ (Figure 6). If $\mathcal{P}$ is close to $\mathcal{X}$, then the average distance between $\mathcal{P}$ and the set of $\mathcal{Y}_2$ is close to the average distance of $\mathcal{X}$ and $\mathcal{Y}_1$. Unlike the traditional MNSQ fit tests, this time one does not only have the average values, but both distributions are present making realistic comparison possible.

An awkward feature of this test is that 1 has to be subtracted from one of the distributions. 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 6 to 9. This topic was detailed previously in Theorem 1.

4.6 Fit Test MC$_{\mathcal{Y}\mathcal{X}}$

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 between $\mathcal{X}$ and the set of $\mathcal{Y}$ is the same as the average distance among the $\mathcal{Y}$ matrices (Figure 7). Here, again, one has two distributions resulting in two p-values (p and p').

Figure 6: Geometry Behind $MC_{\mathcal{P}\mathcal{X}}$

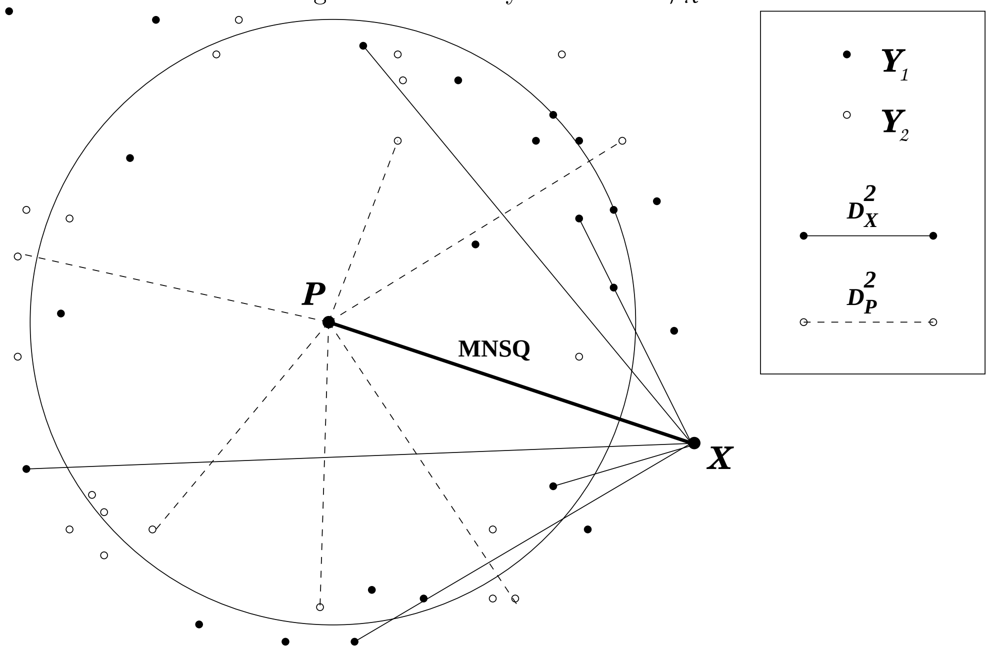


Figure 7: Geometry Behind $MC_{\mathcal{Y}\mathcal{X}}$

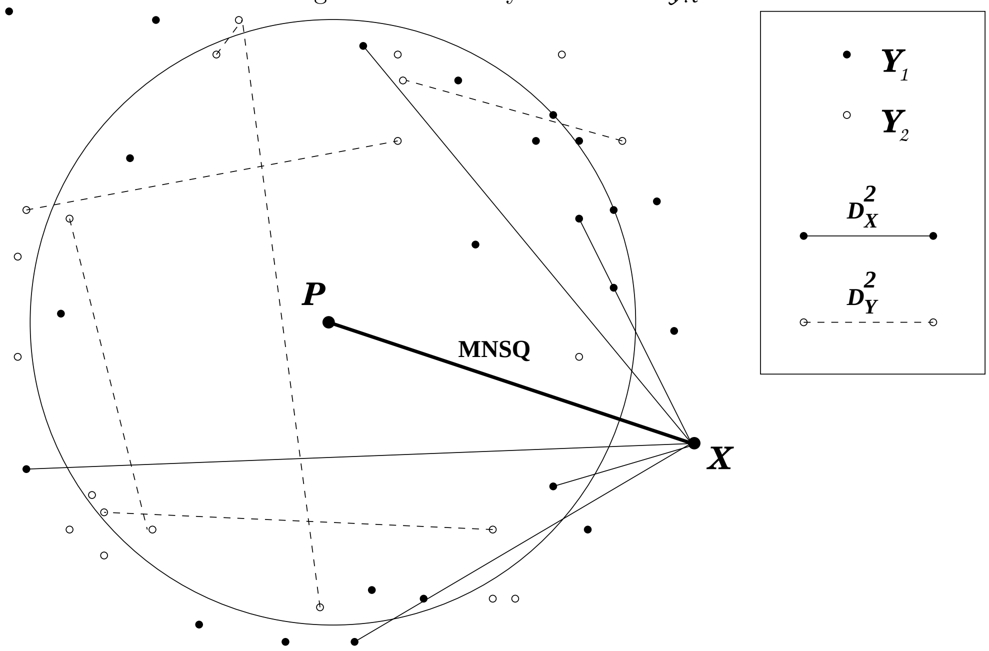
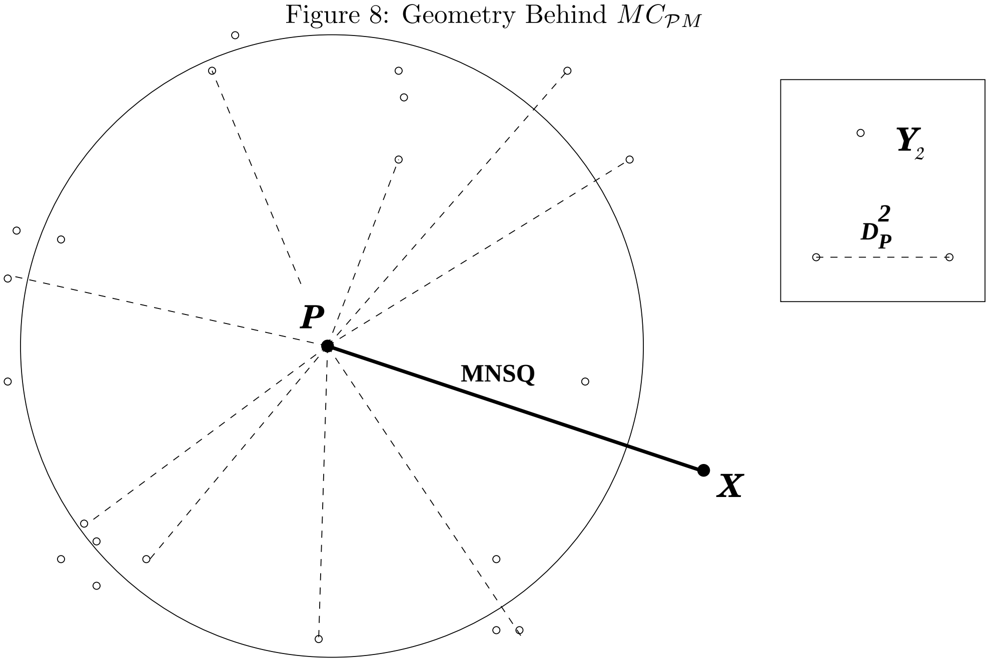


Figure 8: Geometry Behind $MC_{\mathcal{P}M}$



4.7 Fit Test MC$_{\mathcal{P}M}$

The starting point of this test (Figure 8) is that any simulated $\mathcal{Y}$ matrix could be a perfectly acceptable response matrix. That is, the distance squared between $\mathcal{X}$ and $\mathcal{P}$ (MNSQ) should not differ from the mean of the squared distances of $\mathcal{Y}$ and $\mathcal{P}$. When significant deviation is found, misfit is reported.

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.

4.8 Fit Test MC$_{\mathcal{Y}M}$

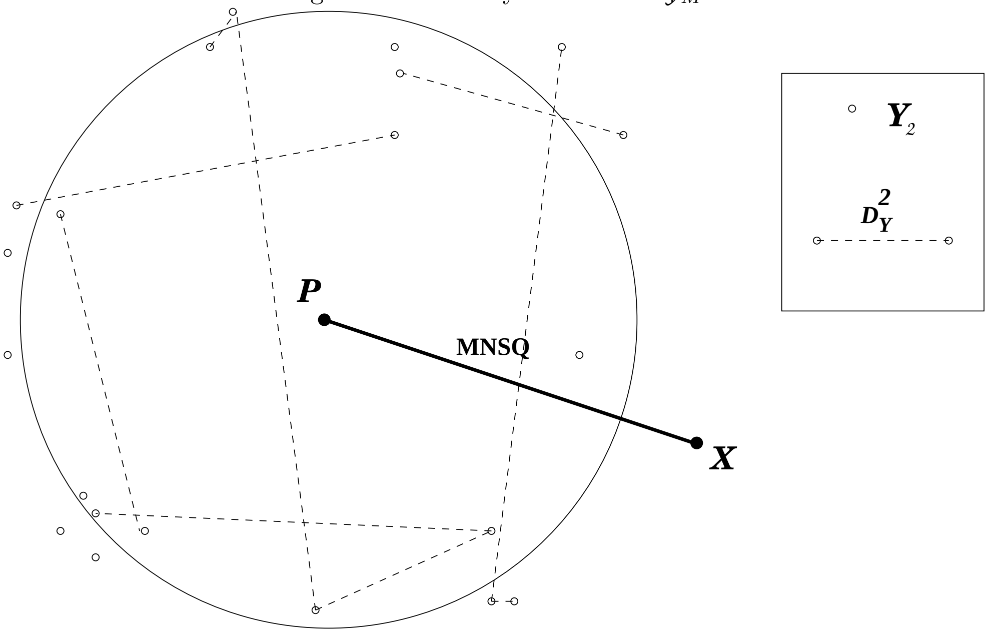
Yet another extension of the ideas above is the test denoted by $MC_{\mathcal{Y}M}$. The underlying geometric picture is explained as follows (Figure 9). To assess the distance between $\mathcal{X}$ and $\mathcal{P}$ first one "blows-up" $\mathcal{P}$ into a set of response matrices and then substitutes $\mathcal{X}$ 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 5).

4.9 Case of General Fit Indices

In this section an indication is given how to extend the above described procedure 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

Figure 9: Geometry Behind $MC_{\mathcal{Y}M}$



its response probability matrix $\mathcal{P}$; the specifics of the actual IRT are encoded in the structure of $\mathcal{P}$. As before, the elements of $\mathcal{P}$ are the probabilities of the actual performance levels given all the model parameters. Any IRT model defines a set of response matrices, that is the set of matrices with size $N \times L$ (N and L are as before) and with entries in accordance with 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 goal 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 particular models lies in the way one may (or may not) 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

$$V = \sum_{i=1}^{L} \sum_{j=1}^{N} (x_{ji} - P_{ji}) w_{ji} \tag{45}$$

or

$$V^* = \sum_{i=1}^{L} \sum_{j=1}^{N} (x_{ji} - P_{ji})^2 v_{ji}, \tag{46}$$

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 45 and 46. 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 45 and 46). For simplicity we only cover $\mathrm{MC}_{\mathcal{P}M}^{V}$.

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

$$\begin{aligned} V_k := V(\mathcal{P}, \mathcal{Y}) &:= \sum_{i=1}^{L} \sum_{j=1}^{N} (y_{ji}^k - P_{ji}) w_{ji}, && (47) \\ V_k^* := V^*(\mathcal{P}, \mathcal{Y}) &:= \sum_{i=1}^{L} \sum_{j=1}^{N} (y_{ji}^k - P_{ji})^2 v_{ji}. && (48) \end{aligned}$$

Then, one performs the hypotheses testing to see if

$$V = V(\mathcal{P}, \mathcal{X}) \quad \text{and} \quad V^* = V^*(\mathcal{P}, \mathcal{X}) \tag{49}$$

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 earlier about the correspondence between the relative frequency of $\mathcal{Y}$ in the simulation scheme advocated here and the probability $\mathfrak{p}(\mathcal{Y})$ of $\mathcal{Y}$ makes it possible to find the p-value of the index in question.

5 Comparison and Stability Studies

5.1 Overview of the Study

Two studies are carried out to investigate certain characteristics of the new method. First, the Monte Carlo fit tests are compared to the traditionally used MNSQ outfit test for response matrices of different size to reveal their association. Then, the stability over the number of simulations is shown.

Original response matrices for these studies were obtained from the Mathematics Department of the Ohio State University. Math 116 (”Excursions in Mathematics”) multiple choice final (48 items) and midterm (24 items) exams, both were administered to 82 students. Sub-tests were created by chopping off approximately half of the group of examinees. Consequently, four different response matrices were obtained for the analyses with respective sizes: $82 \times 48,\ \ 40 \times 48,\ \ 40 \times 24,\ \ 82 \times 24$.

In the following tables (Tables 1-6), the rows of p_{χ^2} show the result of MNSQ outfit (after applying the Wilson-Hilferty transformation and numerical integration). The next five rows show the respective Monte Carlo p-values resulted from five independent replications. The tables are in pairs. The first table is for the case when the number of simulated matrices is $K = 1000$, while the second contains the results from $K = 5000$ simulations.

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

5.2 Summary of Tables

The two charts in Table 1 show the values of $p_{\mathcal{P}M}$. As for the issue of comparison one may record that while p_{χ^2} shows misfit at $\alpha = 0.05$ in two cases (those that are highlighted) $p_{\mathcal{P}M}$ shows moderate fit in all cases. Note, that this observation is yet another indication of the frequently cited lack of normality of MNSQ$_{zstd}$. Even in those cases when p_{χ^2} signals fit $p_{\mathcal{P}M}$ is significantly higher than p_{χ^2}.

Table 1: Values of $p_{\mathcal{P}M}$ for four different tests.

K=1000	82×48	40×48	40×24	82×24	K=5000	82×48	40×48	40×24	82×24
p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**	p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**
$p_{\mathcal{P}M}$	0.3100	0.1810	0.3260	0.2260	$p_{\mathcal{P}M}$	0.3200	0.1818	0.2988	0.2338
$p_{\mathcal{P}M}$	0.3240	0.1740	0.2900	0.2260	$p_{\mathcal{P}M}$	0.3188	0.1840	0.3082	0.2240
$p_{\mathcal{P}M}$	0.2880	0.1800	0.3090	0.2080	$p_{\mathcal{P}M}$	0.3218	0.1804	0.2976	0.2320
$p_{\mathcal{P}M}$	0.3270	0.1620	0.3100	0.2410	$p_{\mathcal{P}M}$	0.3204	0.1742	0.2960	0.2342
$p_{\mathcal{P}M}$	0.3340	0.1770	0.2900	0.2390	$p_{\mathcal{P}M}$	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

Note: Five replications were created to assess stability of the MC indices.

Table 2: Values of $p_{\mathcal{Y}M}$ for four different tests.

K=1000	82×48	40×48	40×24	82×24	K=5000	82×48	40×48	40×24	82×24
p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**	p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**
$p_{\mathcal{Y}M}$	0.3710	0.2640	0.3900	0.3220	$p_{\mathcal{Y}M}$	0.3856	0.2660	0.3652	0.3366
$p_{\mathcal{Y}M}$	0.3820	0.2620	0.3600	0.3220	$p_{\mathcal{Y}M}$	0.3804	0.2772	0.3688	0.3142
$p_{\mathcal{Y}M}$	0.3650	0.2700	0.3820	0.3240	$p_{\mathcal{Y}M}$	0.3786	0.2732	0.3710	0.3244
$p_{\mathcal{Y}M}$	0.3760	0.2590	0.3530	0.3230	$p_{\mathcal{Y}M}$	0.3774	0.2660	0.3648	0.3252
$p_{\mathcal{Y}M}$	0.3810	0.2650	0.3650	0.3360	$p_{\mathcal{Y}M}$	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

Note: Five replications were created to assess stability of the MC indices.

The stability analysis for $p_{\mathcal{P}M}$ shows a very satisfactory result. Even though the range of $p_{\mathcal{P}M}$ reaches 0.0460 (for sample size $K = 1000$; 82×48) it drops below 0.012 for all tests when the sample size is raised to 5000. The means show a 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, when K was raised to 5000 from 1000).

One may conclude that for the test sizes under investigation the Monte Carlo test $p_{\mathcal{P}M}$ with 5000 replication yields reliably stable p-values.

In general, the comparison between the traditional MNSQ and the new Monte Carlo test can be performed easily, as the result is consistently the same for all cases of Monte Carlo indices: the Monte Carlo tests never show significant misfit (not even at the strict $\alpha = 0.1$ level), while MNSQ outfit shows significant (at $\alpha = 0.05$) misfit in two out of four cases.

The question of stability shows an overall remarkable result. Except for $p_{\mathcal{Y}\mathcal{X}}$ and $p'_{\mathcal{Y}\mathcal{X}}$ (which is really the same case) the range of p-values with sample size $K = 5000$ is rarely over 0.02. The means always show very good stability, as well. It is worthwhile to note it once more, that a sample of size 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 reliable result. This is due to the way this humongous set is sampled. Those matrices which are likely to be sampled are those with high probability (with respect to the measure $\mathfrak{p}$). Most of the matrices in this set come with negligible probability, even when compared to the size of the set M_{resp}.

It is easy to understand why $p_{\mathcal{Y}\mathcal{X}}$ and $p'_{\mathcal{Y}\mathcal{X}}$ produce the least stable results (Tables 5 and 6) as the sampled response matrices are used the most intensively in these two tests. These p-values are calculated from two distributions, one of which is the set of distances *among* simulated matrices

Table 3: Values of $p_{\mathcal{P}\mathcal{X}}$ for four different tests.

K=1000	82×48	40×48	40×24	82×24
p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**
$p_{\mathcal{P}\mathcal{X}}$	0.3040	0.2270	0.3010	0.2340
$p_{\mathcal{P}\mathcal{X}}$	0.3280	0.2040	0.3020	0.2320
$p_{\mathcal{P}\mathcal{X}}$	0.2950	0.2360	0.3020	0.2470
$p_{\mathcal{P}\mathcal{X}}$	0.3040	0.2250	0.2940	0.2440
$p_{\mathcal{P}\mathcal{X}}$	0.3090	0.2360	0.3030	0.2740
Mean	0.3080	0.2256	0.3004	0.2462
Std.Dev.	0.0123	0.0131	0.0036	0.0168
Range	0.0330	0.0320	0.0090	0.0420
Minimum	0.2950	0.2040	0.2940	0.2320
Maximum	0.3280	0.2360	0.3030	0.2740

K=5000	82×48	40×48	40×24	82×24
p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**
$p_{\mathcal{P}\mathcal{X}}$	0.2970	0.2194	0.2874	0.2536
$p_{\mathcal{P}\mathcal{X}}$	0.2976	0.2312	0.2924	0.2366
$p_{\mathcal{P}\mathcal{X}}$	0.3120	0.2232	0.2940	0.2584
$p_{\mathcal{P}\mathcal{X}}$	0.3054	0.2244	0.2954	0.2512
$p_{\mathcal{P}\mathcal{X}}$	0.3072	0.2300	0.2898	0.2540
Mean	0.3038	0.2256	0.2918	0.2508
Std.Dev.	0.0064	0.0049	0.0032	0.0083
Range	0.0150	0.0118	0.0080	0.0218
Minimum	0.2970	0.2194	0.2874	0.2366
Maximum	0.3120	0.2312	0.2954	0.2584

Note: Five replications were created to assess stability of the MC indices.

Table 4: Values of $p'_{\mathcal{P}\mathcal{X}}$ for four different tests.

K=1000	82×48	40×48	40×24	82×24
p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**
$p'_{\mathcal{P}\mathcal{X}}$	0.3230	0.1810	0.3420	0.2110
$p'_{\mathcal{P}\mathcal{X}}$	0.3300	0.1740	0.2870	0.2140
$p'_{\mathcal{P}\mathcal{X}}$	0.2650	0.1810	0.3080	0.2130
$p'_{\mathcal{P}\mathcal{X}}$	0.3280	0.1620	0.3240	0.2240
$p'_{\mathcal{P}\mathcal{X}}$	0.3380	0.1810	0.2990	0.2540
Mean	0.3168	0.1758	0.3120	0.2232
Std.Dev.	0.0295	0.0083	0.0215	0.0179
Range	0.0730	0.0190	0.0550	0.0430
Minimum	0.2650	0.1620	0.2870	0.2110
Maximum	0.3380	0.1810	0.3420	0.2540

K=5000	82×48	40×48	40×24	82×24
p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**
$p'_{\mathcal{P}\mathcal{X}}$	0.3174	0.1796	0.3000	0.2358
$p'_{\mathcal{P}\mathcal{X}}$	0.3168	0.1838	0.3098	0.2116
$p'_{\mathcal{P}\mathcal{X}}$	0.3268	0.1810	0.2976	0.2374
$p'_{\mathcal{P}\mathcal{X}}$	0.3178	0.1792	0.2982	0.2348
$p'_{\mathcal{P}\mathcal{X}}$	0.3162	0.1754	0.2938	0.2314
Mean	0.3190	0.1798	0.2999	0.2302
Std.Dev.	0.0044	0.0030	0.0060	0.0106
Range	0.0106	0.0084	0.0160	0.0258
Minimum	0.3162	0.1754	0.2938	0.2116
Maximum	0.3268	0.1838	0.3098	0.2374

Note: Five replications were created to assess stability of the MC indices.

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 0.0314 (when $K = 5000$, test 82×24). Since the p-values themselves are all larger than 0.27, this deviation is relatively small, and in no way can affect the conclusion of the hypothesis tests.

The only case when the small inaccuracy of the Monte Carlo tests can have effect on the final conclusion is when the p-value is around the prescribed significance level α. An extensive simulation study (Antal & Antal, 2003a) shows, however, that this is almost never the case. That is, the p-value of a response matrix is very unlikely to indicate significant misfit! The above cited study could not exhibit a single response matrix $\mathcal{X}$ showing misfit at $\alpha = 0.01$ and just a couple showed significant misfit at $\alpha = 0.05$ (out of 100,000 carefully chosen response matrices).

6 Conclusion

The non-parametric family of Monte Carlo fit tests introduced in this paper 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 p-values.

None of the tests showed significant misfit for the response matrices under consideration, contradicting the traditional MNSQ outfit test. At the same time this shows, that the χ^2 (or normality) assumption of MNSQ is not appropriate, an old result which actually initiated this research.

Although the Monte Carlo p-values are always approximate, they show remarkable stability

Table 5: Values of $p_{\mathcal{YX}}$ for four different tests.

K=1000	82×48	40×48	40×24	82×24	K=5000	82×48	40×48	40×24	82×24
p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**	p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**
$p_{\mathcal{YX}}$	0.2960	0.2250	0.2910	0.2410	$p_{\mathcal{YX}}$	0.2986	0.2192	0.2802	0.2536
$p_{\mathcal{YX}}$	0.3220	0.2080	0.2900	0.2320	$p_{\mathcal{YX}}$	0.2996	0.2306	0.2836	0.2402
$p_{\mathcal{YX}}$	0.2990	0.2360	0.3160	0.2520	$p_{\mathcal{YX}}$	0.3106	0.2226	0.2920	0.2612
$p_{\mathcal{YX}}$	0.2980	0.2250	0.2800	0.2560	$p_{\mathcal{YX}}$	0.3060	0.2178	0.2922	0.2494
$p_{\mathcal{YX}}$	0.3000	0.2300	0.3030	0.2690	$p_{\mathcal{YX}}$	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

Note: Five replications were created to assess stability of the MC indices.

Table 6: Values of $p'_{\mathcal{YX}}$ for four different tests.

K=1000	82×48	40×48	40×24	82×24	K=5000	82×48	40×48	40×24	82×24
p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**	p_{χ^2}	0.1021	**0.0145**	0.0914	**0.0447**
$p'_{\mathcal{YX}}$	0.3750	0.2660	0.3960	0.3120	$p'_{\mathcal{YX}}$	0.3842	0.2644	0.3654	0.3386
$p'_{\mathcal{YX}}$	0.3840	0.2610	0.3590	0.3190	$p'_{\mathcal{YX}}$	0.3792	0.2760	0.3696	0.3072
$p'_{\mathcal{YX}}$	0.3530	0.2700	0.3820	0.3260	$p'_{\mathcal{YX}}$	0.3816	0.2748	0.3708	0.3292
$p'_{\mathcal{YX}}$	0.3770	0.2590	0.3590	0.3100	$p'_{\mathcal{YX}}$	0.3752	0.2698	0.3658	0.3256
$p'_{\mathcal{YX}}$	0.3850	0.2730	0.3730	0.3480	$p'_{\mathcal{YX}}$	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

Note: Five replications were created to assess stability of the MC indices.

with respect to the number of simulated matrices. It was shown that for moderately sized response matrices a sample of size 5000 gives satisfactory result. It is always a good idea, however, to 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 reliable result. Some replications of the calculation of the Monte Carlo p-values, however, could easily lead to reasonably well founded decision about fit. In this study 5 replications were made for both sample sizes $K = 1000$ and $K = 5000$.

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 was supposed to display significant misfit frequently (and misleadingly), the increased computational time should be worth clarifying the goodness fit of a particular model. As high computing power becomes more and more commonplace there is nothing to obstruct the way of serious fit analysis.

References

[1] Antal, J. (2003). A person fit test based on Monte-Carlo method. Paper presentation at the annual meeting of the National Council on Measurement in Education, Chicago, IL

[2] Antal, T. & Antal, J. (2003a). Global Rasch fit analysis. (unpublished manuscript)

[3] Antal, T. & Antal, J. (2003b). *Kardinál* 0.018: Comprehensive Rasch Analysis. Computer Software.

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

[5] Baker, F. (2001). *The basics of Item Response Theory.* ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.

[6] Chen, W.H., & Thissen, D. (1999). Estimation of item parameters for the three-parameter logistic model using the marginal likelihood of summed scores. *British Journal of Mathematical and Statistical Psychology, 52*, 19-37.

[7] Cohen, A. S. (1991). Influence of prior distributions on detection of DIF. *Journal of Educational Measurement, 28,* 1, 49-59.

[8] 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.

[9] Ferrara, S. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large-scale performance assessment. *Applied Measurement in Education, 10,* 2, 123-44.

[10] Ferrara, S.; Huynh, H. & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Measurement, 36,* 2, 119-40.

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

[12] 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.

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

[14] 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.

[15] Linacre, J. M. & Wright, B.D. (2002). Understanding Rasch measurement: construction of measures from many-facet data. *Journal of Applied Measurement 3,* 4, p486-512.

[16] Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement. 25,* 2, 107-135.

[17] 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.

[18] Parshall, C. G. & Miller, T. R.(1995). Exact versus asymptotic Mantel-Haenszel DIF Statistics: a comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32,* 3, 302-316.

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

[20] Rao, A. R.; Jana, R. and Bandyopadhyay, S. (1996). A Markov chain Monte Carlo method for generating random $(0,1)$ matrices with given marginals. *Sankhya ser.* A 58, 225-242.

[21] Reese. L. M. (1999). Impact of local item dependence on item response theory scoring in CAT. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

[22] Roberts, J. M. Jr. (2000). Simple methods for simulating sociomatrices with given marginal totals. *Social Networks* 22, 273-283.

[23] Roussos, L. A., Schnipke, D. L. & Pashley, P.J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24,*3 293-322.

[24] Smith, E. V. Jr. (2001). Understanding Rasch measurement: Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. *Journal of Applied Measurement, 2*, 3, 281-311.

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

[26] Snijders, T. A. B. (1991). Enumeration and simulation methods for 0-1 matrices with given marginals. *Psychometrika*, 56, 397-417

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

Stone, G. (2001). Understanding Rasch measurement: Objective standard setting (or truth in advertising). *Journal of Applied Measurement, 2,*2, 187-201.

[28] Swaminathan, H. & Rogers, H. J.(1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,*4, 361-70.

[29] Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement, 12,* 4, 353-364.

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

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

[32] Wright, B. D. & Mok, M. (2000). Understanding Rasch Measurement: Rasch Models Overview. *Journal of Applied Measurement, 1,* 1, 83-103.

[33] Wright, B. D. and Stone, M. H (1979). *Best test design.* Chicago: MESA Press.

[34] Yen, W. M. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,*3, 187-213.

[35] Zwick R., Thayer D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel differential item functioning analysis to a computerized adaptive test. *Applied Psychological Measurement, 26,* 1, 57-76.