PSYCHOMETRIKA— VOL. 70, NO. 3, 497–516
SEPTEMBER 2005
DOI: 10.1007/s11336-002-1040-6

CONSTANT LATENT ODDS-RATIOS MODELS AND THE MANTEL-HAENSZEL NULL HYPOTHESIS

DAVID J. HESSEN

UNIVERSITY OF AMSTERDAM

In the present paper, a new family of item response theory (IRT) models for dichotomous item scores is proposed. Two basic assumptions define the most general model of this family. The first assumption is local independence of the item scores given a unidimensional latent trait. The second assumption is that the odds-ratios for all item-pairs are constant functions of the latent trait. Since the latter assumption is characteristic of the whole family, the models are called constant latent odds-ratios (CLORs) models. One nonparametric special case and three parametric special cases of the general CLORs model are shown to be generalizations of the one-parameter logistic Rasch model. For all CLORs models, the total score (the unweighted sum of the item scores) is shown to be a sufficient statistic for the latent trait. In addition, conditions under the general CLORs model are studied for the investigation of differential item functioning (DIF) by means of the Mantel-Haenszel procedure.

Key words: IRT, odds-ratio, sufficient statistic, weak item independence, manifest monotonicity, Rasch model, DIF, Mantel-Haenszel procedure.

1. Introduction

Fundamental to item response theory (IRT) models for dichotomous item scores is the item response function (IRF). The IRF relates the probability of a correct or positive item score to a latent trait that is assumed to underlie the performance on a set of test items. A distinction can be made between parametric and nonparametric IRT models. In parametric IRT models, the IRF is defined by a prespecified mathematical function, whereas in nonparametric IRT models the IRF is defined by certain order restrictions. Well-known examples of both types of models are the one-, two-, and three-parameter logistic models (Rasch, 1960; Birnbaum, 1968) and the nonparametric monotone homogeneity and double monotonicity models (Mokken, 1971; Mokken and Lewis, 1982; Mokken, Lewis, and Sijtsma, 1986).

A well-known result for the two-parameter logistic model (Birnbaum, 1968) is that the weighted sum of the item scores is a sufficient statistic for the latent trait, where the weights are the discrimination parameters. From this result, it follows for the one-parameter logistic model (Rasch, 1960) that the total score (the unweighted sum of the item scores) is a sufficient statistic for the latent trait. In addition, Fischer (1974, 1995a) showed that the following assumptions imply a model equivalent to the one-parameter logistic Rasch model: (1) continuous, strictly monotone increasing IRFs, (2) all IRFs with lower limit zero and upper limit one, (3) local independence (LI) of the item scores given a unidimensional latent trait, and (4) sufficiency of the total score for the latent trait.

In the present paper, it is shown that there exist other IRT models than the Rasch model under which the total score is a sufficient statistic for the latent trait when both the assumptions of continuous, strictly monotone IRFs and that of all IRFs with lower limit zero and a upper limit

This research was supported by the Dutch Organization for Scientific Research (NWO), grant number 400-20-026.

Requests for reprints should be sent to David J. Hessen, Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, THE NETHERLANDS. E-Mail: D.J.Hessen@uva.nl

© 2005 The Psychometric Society

one are dropped. In order to show that there exist other models with the sufficiency property of the total score, it is convenient to focus on the positions of the IRFs relative to one another or on the interrelationships between the IRFs, instead of focusing on the shape and/or the location of the IRF of each separate item. A useful aid to the comparison of two IRFs is the odds-ratio function (ORF) (Rosenbaum, 1987). The ORF is the ratio of the odds of a correct item score between two items as a function of the latent trait. If the only additional assumption to local independence of the item scores is the assumption that the ORF for each pair of items is a constant function of the latent trait, then a more general measurement model than the one-parameter logistic Rasch model is obtained, which is called the general constant latent odds-ratios (CLORs) model.

In the framework of IRT, the sufficiency of an observed score (such as the total score) for the latent trait is an important measurement property for the study of differential item functioning (DIF). Under IRT models for dichotomous item scores, DIF is defined to exist for a particular item when its IRFs in at least two groups of examinees are unequal. Items must be tested for DIF to prevent the measurement of unintended factors or variables and to guarantee fair testing of examinees from different demographic or biographic groups. For this purpose, observed score-based DIF detection methods can be considered (Millsap and Everson, 1993). Such methods make use of an observed matching variable as a proxy for the latent trait. However, Meredith and Millsap (1992) have shown that DIF detection methods that rely exclusively on an observed matching variable are not generally diagnostic of DIF. They have shown that in theory these methods can only be diagnostic of DIF in case at least the observed matching variable is a sufficient statistic for the latent trait.

One of the standard procedures for the detection of DIF in dichotomous item scores, which uses an observed score as a proxy for the latent trait, is the Mantel–Haenszel (MH) procedure (Mantel and Haenszel, 1959; Holland and Thayer, 1988). In the MH procedure, a Chi-square test statistic with one degree of freedom is used to test the null hypothesis whether the odds of giving a correct response to a so-called 'studied item' are equal in two different groups at each level of the selected observed matching criterion. It has been shown that this MH null hypothesis and the IRT definition of no DIF do not coincide in general (Holland and Thayer, 1988; Zwick, 1990). Up to now, the MH null hypothesis and the IRT definition of no DIF have only been shown to be equal under the following restrictive conditions (Holland and Thayer, 1988; Fischer, 1993, 1995b): (1) the data follow the one-parameter logistic Rasch model, (2) the criterion for matching includes the studied item, (3) none of the items in the criterion for matching, except for possibly the studied item, is a DIF item, and (4) the data of the two groups are random samples from the corresponding populations. In this paper, the conditions under which the MH procedure is theoretically appropriate for DIF detection under the general CLORs model are studied. Manifest odds-ratios used in the MH procedure are derived in terms of CLORs and show under which circumstances the MH null hypothesis is met and under which it is equal to the IRT definition of no DIF.

First, in the next section, the general model with CLORs for all item-pairs is discussed. It is proved that under this model the total score is a sufficient statistic for the latent trait. In addition, two simple observable consequences are derived that can be used in evaluating the model. Subsequently, one nonparametric and three parametric special cases of the general CLORs model are discussed and are shown to be unequal to the one-parameter logistic Rasch model. In the following section, the manifest MH odds-ratios are derived in terms of CLORs. Under the general CLORs model, it is discussed how the MH null hypothesis is related to the IRT definition of no DIF. Then, by means of an example a so-called worst case scenario is illustrated in which the MH null hypothesis is satisfied and DIF is present for all items.

2. A Family of CLORs Models

2.1. Assumptions of the General CLORs Model

The general model under consideration is a model for dichotomous items that are assumed to measure the same unidimensional latent trait. Let T be a test consisting of k dichotomous items. The item score variable for the response of an examinee to item i is denoted by X_i and

$$X_i = \begin{cases} 1, & \text{if the response of an examinee to item } i \text{ is correct or positive} \\ 0, & \text{if the response is incorrect or negative,} \end{cases} \tag{1}$$

for $i = 1, \ldots, k$. The vector of item score variables is denoted by $\mathbf{X} = (X_1, \ldots, X_k)$. The vector of item score realizations is denoted by $\mathbf{x} = (x_1, \ldots, x_k)$. Let $P(\mathbf{X} = \mathbf{x})$ denote the joint probability function of response pattern $\mathbf{x}$ in population C if tested by test T. By letting $\mathbf{x}$ range over its 2^k possible values, the set $\{P(\mathbf{X} = \mathbf{x})\}$ is formed, which constitutes the manifest probabilities for the test T in population C. Then,

$$P(\mathbf{X} = \mathbf{x}) \geq 0, \quad \text{for all } \mathbf{x} \quad \text{and} \quad \sum_{x} P(\mathbf{X} = \mathbf{x}) = 1.$$

Unidimensional IRT models specify a functional form for the manifest probabilities by assuming that there is a unidimensional latent trait or ability θ and that $P(\mathbf{X} = \mathbf{x})$ has the representation:

$$P(\mathbf{X} = \mathbf{x}) = \int \prod_{i=1}^{k} P_i(\theta)^{x_i} Q_i(\theta)^{1-x_i} \, \mathrm{d}F(\theta), \quad \text{for all } \mathbf{x}, \tag{2}$$

where $P_i(\theta)$ is the IRF for item i, $Q_i(\theta) = 1 - P_i(\theta)$, and $F(\theta)$ is the cumulative distribution function of the latent trait θ over the population C. The structure in Equation 2 includes the assumption of LI of the item scores (Lord and Novick, 1968), defined as

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \prod_{i=1}^{k} P_i(\theta)^{x_i} Q_i(\theta)^{1-x_i}, \quad \text{for all } \mathbf{x} \text{ and all } \theta, \tag{3}$$

and means that the item scores are conditionally independent given the latent trait. Rewriting the definition of LI in Equation 3 in terms of the item passing-odds curves (IPCs), defined as $V_i(\theta) = \frac{P_i(\theta)}{Q_i(\theta)}$, for $i = 1, \ldots, k$ (Cressie and Holland, 1983), yields the slightly more compact version

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \prod_{i=1}^{k} V_i(\theta)^{x_i} Q_i(\theta), \quad \text{for all } \mathbf{x} \text{ and all } \theta. \tag{4}$$

Furthermore, it is assumed that the probability of giving a correct response to each item conditional on the latent trait is never equal to zero or one; that is, $0 < P_i(\theta) < 1$, for all i. At this point, it is stressed that under the general model to be discussed in the first place none of the IRFs is restricted to the limit assumptions

$$\lim_{\theta \to -\infty} P_i(\theta) = 0 \quad \text{and} \quad \lim_{\theta \to \infty} P_i(\theta) = 1,$$

which hold for each IRF under the one-parameter logistic Rasch model. So, each IRF is allowed to have a lower asymptote greater than zero and an upper asymptote smaller than one.

The characteristic assumption of the general model considered concerns the relationship between the ratio of the odds for any pair of items and the latent trait. The odds-ratio is a well-known simple measure for the association between two dichotomous variables. Within IRT, the

odds-ratio is just like the probability of a correct response, a function of the latent trait. By using IPCs, the odds-ratio for any two dichotomous items i and j as a function of the latent trait is defined as (Rosenbaum, 1987)

$$\omega_{ij}(\theta) = \frac{V_i(\theta)}{V_j(\theta)}, \quad \text{for all } i \neq j, \tag{5}$$

and is called the ORF for items i and j. In case of k items, there are $\binom{k}{2}$ such ORFs, which are not independent of one another. Each ORF can be obtained by two other ORFs, for example: $\omega_{ik}(\theta) = \omega_{ij}(\theta)\omega_{jk}(\theta) = \frac{V_i(\theta)}{V_j(\theta)}\frac{V_j(\theta)}{V_k(\theta)} = \frac{V_i(\theta)}{V_k(\theta)}$. This set of dependent ORFs can be reduced to a minimal set of $\binom{k}{2} - \binom{k-1}{2} = k-1$ ORFs containing all information. One such set can be obtained by expressing the IPCs of $k-1$ items, denoted $V_i(\theta)$, for $i = 1, \ldots, k-1$, in terms of the IPC of the arbitrary selected, fixed reference item k; that is,

$$V_i(\theta) = \omega_{ik}(\theta)V_k(\theta), \quad \text{for } i = 1, \ldots, k-1, \tag{6}$$

which will be used in the rest of this paper.

The structure in Equation 2 with the assumption of LI is only a representation rather than a measurement model (Rosenbaum, 1987), because it does not itself restrict the manifest joint probability function $P(\mathbf{X} = \mathbf{x})$ in any way. Therefore, additional assumptions are required to place testable restrictions on the observable data. Here, to obtain an IRT model from which observable consequences follow, it is additionally assumed that the $k-1$ ORFs from the minimal set in Equation 6 are constant functions of the latent trait; that is,

$$\omega_{ik}(\theta) = \alpha_{ik}, \quad \text{for } i = 1, \ldots, k-1, \text{ and all } \theta, \tag{7}$$

and will be called the assumption of CLORs for all item-pairs. So far nothing else is assumed in addition to (1) LI of the item scores given a unidimensional latent trait and (2) CLORs for all item-pairs. These two assumptions define the model that will be called the general CLORs model. Without any further assumptions, the IRFs of this general CLORs model are not restricted to be continuous, they may be constant over θ, and they are even allowed to increase and decrease repeatedly with θ. Under the model, only two situations are possible for the lower limits of the IRFs: (1) they are all equal to zero, or (2) none of them equals zero and each of them has a unique value. Likewise, only two situations are possible for the upper limits: (1) they are all equal to one, or (2) none of them equals one and each of them has a unique value.

Next, from the general CLORs model it follows that the ordering of the IRFs is the same for all θ. In other words, the general CLORs model has the property of invariant item ordering (IIO) (Sijtsma and Junker, 1996). The definition of IIO for k numbered and ordered items is

$$P_1(\theta) \geq P_2(\theta) \geq \cdots \geq P_k(\theta), \quad \text{for all } \theta, \tag{8}$$

and means that the IRFs of the set of k items can touch but do not intersect. For example, in Equation 8 item 1 is the easiest item for all examinees and item k is the most difficult item for all examinees.

Proposition 1. If for two items i and j, $\omega_{ij}(\theta) = \alpha_{ij} > 1$, for all θ, then $P_i(\theta) > P_j(\theta)$, for all θ.

Proof. Since by definition $\omega_{ij}(\theta) = \frac{P_i(\theta)\{1-P_j(\theta)\}}{\{1-P_i(\theta)\}P_j(\theta)}$, it follows from $\omega_{ij}(\theta) > 1$, for all θ, that $\frac{P_i(\theta)\{1-P_j(\theta)\}}{\{1-P_i(\theta)\}P_j(\theta)} > 1$, for all θ. Then, $\frac{P_i(\theta)}{1-P_i(\theta)} > \frac{P_j(\theta)}{1-P_j(\theta)}$, for all θ, and some algebra shows that $P_i(\theta) > P_j(\theta)$, for all θ. □

By using the proper logical operators, it can be shown in the same way as for Proposition 1 that if $\alpha_{ij} = 1$, then $P_i(\theta) = P_j(\theta)$, for all θ, and that if $\alpha_{ij} < 1$, then $P_i(\theta) < P_j(\theta)$, for all θ. Under the general CLORs model, these relationships hold for any combination of two items, and thus, the IRFs of all k items do not intersect. The type of IIO under the general CLORs model is however a slightly stronger condition than the definition in Equation 8, because if two IRFs touch, then they must coincide completely, otherwise they are not equal for all θ.

2.2. Sufficiency of the Total Score for the Latent Trait

A popular and often used function of the item scores is the total score, which is defined as $X_+ = \sum_{i=1}^{k} X_i$. The total score is a sufficient statistic for the latent trait if it utilizes all information about the latent trait available from the data. In case the total score is a sufficient statistic for the latent trait, different response patterns with the same total score contain no further information about the latent trait beyond that of the total score.

Theorem 1. Given LI of the item scores, the total score X_+ is a sufficient statistic for θ if and only if the ORFs, denoted by $\omega_{ik}(\theta)$, for $i = 1, \ldots, k-1$, are constants not varying with θ.

Proof. By using the definition of LI in Equation 4 and the minimal set of IPCs in Equation 6, the probability of response pattern $\mathbf{x}$ conditional on θ can be rewritten as

$$\begin{aligned} P(\mathbf{X} = \mathbf{x} \mid \theta) &= \left(\prod_{i=1}^{k} Q_i(\theta)\right)\left(\prod_{i=1}^{k} V_i(\theta)^{x_i}\right) \\ &= \left(\prod_{i=1}^{k} Q_i(\theta)\right)\left(\prod_{i=1}^{k-1} \{\omega_{ik}(\theta) V_k(\theta)\}^{x_i}\right) V_k(\theta)^{x_k} \\ &= \left(\prod_{i=1}^{k} Q_i(\theta)\right)\left(\prod_{i=1}^{k-1} \omega_{ik}(\theta)^{x_i}\right)\left(\prod_{i=1}^{k} V_k(\theta)^{x_i}\right) \\ &= \left(\prod_{i=1}^{k} Q_i(\theta)\right)\left(\prod_{i=1}^{k-1} \omega_{ik}(\theta)^{x_i}\right) V_k(\theta)^{x_+}, \quad \text{for all } \mathbf{x} \text{ and all } \theta. \end{aligned} \tag{9}$$

Subsequently, replacing $\omega_{ik}(\theta)$ by α_{ik}, for $i = 1, \ldots, k-1$, yields

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \left(\prod_{i=1}^{k} Q_i(\theta)\right)\left(\prod_{i=1}^{k-1} \alpha_{ik}^{x_i}\right) V_k(\theta)^{x_+}, \quad \text{for all } \mathbf{x} \text{ and all } \theta, \tag{10}$$

and sufficiency follows by the Neyman Factorization Theorem. This proves the 'if' part of Theorem 1.

Fischer (1974, 1995a) has proved the 'only if' part of Theorem 1. The proof goes as follows. If the total score is a sufficient statistic for the latent trait, then the conditional distribution $P(\mathbf{X} = \mathbf{x} \mid X_+ = x_+, \theta)$, which is equal to $\frac{P(\mathbf{X}=\mathbf{x}|\theta)}{P(X_+=x_+|\theta)}$, is a constant function of the latent trait. Let R_{x_+} be the set of all response patterns for which $X_+ = x_+$; that is, $R_{x_+} = \{\mathbf{x}; \sum_{i=1}^{k} x_i = x_+\}$. Then,

$$P(X_+ = x_+ \mid \theta) = \left(\prod_{i=1}^{k} Q_i(\theta)\right)\left(\sum_{R_{x_+}} \prod_{i=1}^{k-1} \omega_{ik}(\theta)^{x_i}\right) V_k(\theta)^{x_+} \tag{11}$$

and

$$P(\mathbf{X}=\mathbf{x} \mid X_+=x_+,\theta)=\frac{\prod_{i=1}^{k-1}\omega_{ik}(\theta)^{x_i}}{\sum_{R_{x_+}}\prod_{i=1}^{k-1}\omega_{ik}(\theta)^{x_i}}.$$

Let S_{x_+-1} be the set of all respons patterns for which $\sum_{j\neq i}^{k-1}X_j=x_+-1$; that is, $S_{x_+-1}=\{\mathbf{x};\sum_{j\neq i}^{k-1}x_j=x_+-1\}$. Then, if $P(\mathbf{X}=\mathbf{x}\mid X_+=x_+,\theta)$ is a constant function of θ, both the marginal probability

$$P(X_i=x_i,X_k=x_k\mid X_+=x_+,\theta)=\frac{\omega_{ik}(\theta)^{x_i}\sum_{S_{x_+-1}}\prod_{j\neq i}^{k-1}\omega_{jk}(\theta)^{x_j}}{\sum_{R_{x_+}}\prod_{i=1}^{k-1}\omega_{ik}(\theta)^{x_i}},\quad \text{for } i=1,\ldots,k-1,$$

and the probability ratio

$$\frac{P(X_i=1,X_k=0\mid X_+=x_+,\theta)}{P(X_i=0,X_k=1\mid X_+=x_+,\theta)}=\omega_{ik}(\theta),\quad \text{for all } x_+ \text{ and } i=1,\ldots,k-1, \tag{12}$$

are constant functions of θ. □

In what now follows, some simple observable consequences are derived from the general CLORs model. On these observable consequences, methods can be based that allow a rough assessment of whether the general CLORs model can represent the data structure.

Theorem 2. If a set of k items follows the general CLORs model with the property of IIO across θ, then the item-total regressions $P(X_i=1\mid X_+=x_+)$, for $i=1,\ldots,k$, can be ordered such that

$$P(X_1=1\mid X_+=x_+)\geq P(X_2=1\mid X_+=x_+)\geq\cdots\geq P(X_k=1\mid X_+=x_+),\ \text{for all } x_+. \tag{13}$$

Proof. From Equation 12, it follows for any two items i and j that

$$P(X_i=1,X_j=0\mid X_+=x_+)=P(X_i=0,X_j=1\mid X_+=x_+)\alpha_{ij},\ \text{for all } x_+.$$

If $\alpha_{ij}\geq 1$, then

$$P(X_i=1,X_j=0\mid X_+=x_+)\geq P(X_i=0,X_j=1\mid X_+=x_+),\ \text{for all } x_+.$$

Adding $P(X_i=1,X_j=1\mid X_+=x_+)$ to both sides of the inequality yields

$$P(X_i=1\mid X_+=x_+)\geq P(X_j=1\mid X_+=x_+),\quad \text{for all } x_+.$$

Successively repeating these steps for pairs of items that are in a consecutive order yields Equation 13. □

The observable consequence of the general CLORs model in Theorem 2 is called weak item independence (WII) (Scheiblechner, 1995). WII means that the items are ordered isotonically by the conditional response probabilities according to easiness, independent of x_+.

The observable consequence that is implied by the general CLORs model in the following theorem is called manifest monotonicity (Junker, 1993; Junker and Sijtsma, 2000) or weak subject independence (WSI) (Scheiblechner, 1995). The result is somehow surprising because it is still only assumed that the item scores are locally independent and that the ORFs for all item-pairs are constant functions of θ. So far nothing is assumed about the shape of the IRFs.

Theorem 3. If a set of k items follows the general CLORs model, then item-total regressions increase with the total score, that is,

$$P(X_i = 1 \mid X_+ = x_+) \leq P(X_i = 1 \mid X_+ = x_+ + 1), \quad \text{for all } i \text{ and } x_+ = 0, \ldots, k-1. \quad (14)$$

Note that since $P(X_i = 1 \mid X_+ = 0) = 0$ and $P(X_i = 1 \mid X_+ = k) = 1$, for all i, the inequality in (14) for item i, $i = 1, \ldots, k$, only needs to be evaluated for $x_+ = 1, \ldots, k-2$.

Proof. By using Equation 7, the conditional probability distribution of the total score given θ in Equation 11 can be rewritten as

$$P(X_+ = x_+ \mid \theta) = V_k(\theta)^{x_+} \left(\prod_{i=1}^{k} Q_i(\theta) \right) \phi_{x_+}(\boldsymbol{\alpha}), \quad (15)$$

where $\boldsymbol{\alpha} = (\alpha_{1k}, \ldots, \alpha_{(k-1)k}, \alpha_{kk})$ and

$$\phi_{x_+}(\boldsymbol{\alpha}) = \sum_{R_{x_+}} \prod_{i=1}^{k-1} \alpha_{ik}^{x_i} \alpha_{kk}^{x_k}$$

is the elementary symmetric function of total score x_+. Note that $\alpha_{kk}^0 = \alpha_{kk}^1 = 1$. However, in what follows it is notationally convenient to use α_{kk}. The elementary symmetric functions for $x_+ = 0, 1, \ldots, k$, are

$$\begin{aligned}
\phi_0(\boldsymbol{\alpha}) &= 1, \\
\phi_1(\boldsymbol{\alpha}) &= \alpha_{1k} + \alpha_{2k} + \cdots + \alpha_{kk}, \\
\phi_2(\boldsymbol{\alpha}) &= \alpha_{1k}\alpha_{2k} + \alpha_{1k}\alpha_{3k} + \cdots + \alpha_{(k-1)k}\alpha_{kk}, \\
\vdots \quad &= \quad \vdots \\
\phi_k(\boldsymbol{\alpha}) &= \alpha_{1k}\alpha_{2k} \cdots \alpha_{(k-1)k}\alpha_{kk}.
\end{aligned}$$

The following recursion holds

$$\phi_{x_+}(\boldsymbol{\alpha}) = \alpha_{ik}\phi_{x_+-1}(\boldsymbol{\alpha}^{(i)}) + \phi_{x_+}(\boldsymbol{\alpha}^{(i)}),$$

where $\boldsymbol{\alpha}^{(i)}$ is $\boldsymbol{\alpha}$ without α_{ik}. Summing the recursion over i gives

$$k\phi_{x_+}(\boldsymbol{\alpha}) = \sum_{i=1}^{k} \alpha_{ik}\phi_{x_+-1}(\boldsymbol{\alpha}^{(i)}) + \sum_{i=1}^{k} \phi_{x_+}(\boldsymbol{\alpha}^{(i)}).$$

The second sum can be shown to be equal to $(k - x_+)\phi_{x_+}(\boldsymbol{\alpha})$. Therefore,

$$\frac{\sum_{i=1}^{k} \alpha_{ik}\phi_{x_+-1}(\boldsymbol{\alpha}^{(i)})}{\phi_{x_+}(\boldsymbol{\alpha})} = x_+,$$

which is equal to $\sum_{i=1}^{k} P(X_i = 1 \mid X_+ = x_+)$. Since also

$$\frac{\sum_{i=1}^{k} \alpha_{ik}\phi_{x_+}(\boldsymbol{\alpha}^{(i)})}{\phi_{x_++1}(\boldsymbol{\alpha})} = x_+ + 1,$$

it follows that

$$\frac{\sum_{i=1}^{k} \alpha_{ik} \phi_{x_+ - 1}(\boldsymbol{\alpha}^{(i)})}{\phi_{x_+}(\boldsymbol{\alpha})} \leq \frac{\sum_{i=1}^{k} \alpha_{ik} \phi_{x_+}(\boldsymbol{\alpha}^{(i)})}{\phi_{x_+ + 1}(\boldsymbol{\alpha})}.$$

When we proceed in a similar way as Scheiblechner (1995) in proving his Theorem 4, we obtain

$$\frac{\phi_{x_+ - 1}(\boldsymbol{\alpha}^{(i)})}{\phi_{x_+}(\boldsymbol{\alpha})} \leq \frac{\phi_{x_+}(\boldsymbol{\alpha}^{(i)})}{\phi_{x_+ + 1}(\boldsymbol{\alpha})}, \quad \text{for all } i \text{ and } x_+ = 0, 1, \ldots, k-1.$$

Multiplying both sides with α_{ik} gives

$$\frac{\alpha_{ik} \phi_{x_+ - 1}(\boldsymbol{\alpha}^{(i)})}{\phi_{x_+}(\boldsymbol{\alpha})} \leq \frac{\alpha_{ik} \phi_{x_+}(\boldsymbol{\alpha}^{(i)})}{\phi_{x_+ + 1}(\boldsymbol{\alpha})}, \quad \text{for all } i \text{ and } x_+ = 0, 1, \ldots, k-1,$$

which is equal to the inequality in (14). □

So when manifest monotonicity is found in practice, the IRFs can be bell-shaped or single-peaked because the assumptions of LI and CLORs for all item-pairs together do not in any way restrict the general shape that the IRFs have in common.

2.3. A Nonparametric Special Case

An additional assumption imposed on the IRFs of the general CLORs model that yields a nonparametric ordinal CLORs model is the assumption of monotone nondecreasing IRFs or in other words, the assumption of latent monotonicity (Mokken, 1971; Mokken and Lewis, 1982); that is, for two arbitrarily chosen fixed values θ_a and θ_b,

$$P_i(\theta_a) \leq P_i(\theta_b), \quad \text{whenever } \theta_a < \theta_b, \text{ for all } i = 1, \ldots, k. \tag{16}$$

Note that the assumption of nondecreasing IRFs allows the probability of a correct or positive item score to be constant across certain latent trait intervals. The nonparametric ordinal special case of the general CLORs model, which is defined by the assumptions of (1) LI of the item scores given a unidimensional latent trait, (2) CLORs for all item-pairs, and (3) monotone nondecreasing IRFs, is called the monotone CLORs model.

Theorem 4. The monotone CLORs model is a special case of the nonparametric double monotonicity model (Mokken, 1971).

The proof of Theorem 4 follows from the three assumptions that define the double monotonicity model (Mokken, 1971; Mokken and Lewis, 1982). These three assumptions are: (1) LI of the item scores, (2) monotone nondecreasing IRFs, and (3) IIO (Sijtsma and Junker, 1996). The three assumptions of the double monotonicity model specify conditions that also hold for the monotone CLORs model.

Proof. The double monotonicity model and the monotone CLORs model share the assumptions of LI and monotonicity. The property of IIO follows from the assumption of CLORs for all item-pairs (see Proposition 1). □

Since the monotone CLORs model is a special case of the nonparametric double monotonicity model, which in turn is a special case of the monotone homogeneity model (Mokken, 1971), the total score has a monotone likelihood ratio (MLR) in θ (Grayson, 1988; Huynh, 1994).

An important type of stochastic ordering that follows from this MLR of the total score in θ is stochastic ordering of the latent trait by the total score (Junker, 1993; Lehmann, 1994; Hemker, Sijtsma, Molenaar, and Junker, 1997). This stochastic ordering property justifies the use of the total score to order persons on the latent trait.

A graphical example of a set of nine IRFs under the nonparametric monotone CLORs model is given in Figure 1.

2.4. Parametric Special Cases

Under the assumptions of continuous strictly increasing IRFs, LI of the item scores and $\lim_{\theta\to-\infty} P_i(\theta) = 0$ and $\lim_{\theta\to\infty} P_i(\theta) = 1$, for all i, Fischer (1974, 1995a) showed that the one-parameter logistic Rasch model is the only possible model for which X_+ is a sufficient statistic for θ. In what follows, parametric CLORs models that are not equivalent to the one-parameter logistic Rasch model are discussed.

We start out with a special case of the four-parameter logistic model (Barton and Lord, 1981). In the four-parameter logistic model, the lower limits of the IRFs are not necessarily equal to zero and the upper limits are not necessarily equal to one. The four-parameter logistic IRF is given by

$$P_i(\theta) = \frac{c_i + g_i e^{a_i(\theta-b_i)}}{1 + e^{a_i(\theta-b_i)}}, \quad \text{for item } i,\ i = 1, \ldots, k, \tag{17}$$

where $0 \leq c_i < g_i \leq 1$, $a_i > 0$ and $-\infty < b_i < \infty$, for all i. In this four-parameter logistic model, $\lim_{\theta\to-\infty} P_i(\theta) = c_i$ is the lower asymptote, $\lim_{\theta\to\infty} P_i(\theta) = g_i$ is the upper asymptote, b_i is the location parameter, and a_i is the slope parameter of the IRF of item i, $i = 1, \ldots, k$. The c- and g-parameters can be interpreted as parameters that account for guessing and slipping, respectively. Usually, the b- and a-parameters are interpreted as a difficulty and a discrimination parameter, respectively.

Theorem 5. The assumption of LI and a set of k four-parameter logistic IRFs define a special case of the monotone CLORs model, if $a_i = a = 1$, $c_i = \frac{(1-\gamma e^{b_i})\delta}{(1-\delta\gamma)e^{b_i}}$, and $g_i = \frac{e^{b_i}}{\delta} c_i$, for all i, $i = 1, \ldots, k$, where δ and γ are constants greater than or equal to zero.

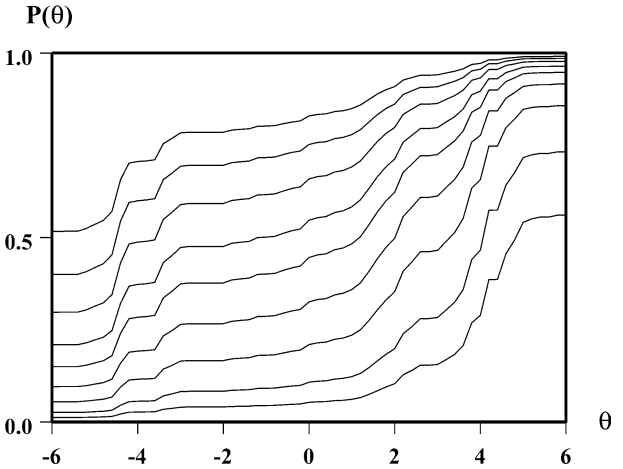

FIGURE 1.
Nine IRFs under the nonparametric monotone CLORs model.

Applying the restrictions $a_i = a = 1$, $c_i = \frac{(1-\gamma e^{b_i})\delta}{(1-\delta\gamma)e^{b_i}}$, and $g_i = \frac{e^{b_i}}{\delta}c_i$, for item $i, i = 1, \ldots, k$, to the four-parameter logistic IRF in Equation 17 yields (Hessen, 2004)

$$P_i(\theta) = \frac{(1-\gamma e^{b_i})(\delta + e^{\theta})}{(1-\delta\gamma)(e^{b_i} + e^{\theta})}, \quad \text{for item } i, i = 1, \ldots, k, \tag{18}$$

where $0 \leq \delta < e^{b_i}$ and $0 \leq \gamma < e^{-b_i}$, for all i. Consequently, $\ln \delta < b_i < -\ln \gamma$, for all i, which means that the range of possible b-parameter values is restricted when $\delta \neq 0$ and/or $\gamma \neq 0$. If $\delta = 0$ and/or $\gamma = 0$, then one of the two range restrictions or both are relaxed, because then $\ln \delta$ or $-\ln \gamma$ or both are not defined.

The model defined by the assumption of LI of the item scores and the IRF in Equation 18 will be called the general parametric CLORs model. In order to show that this model is a special case of the monotone CLORs model, it must be demonstrated that the ORF for any pair of items is a constant function of the latent trait.

Proof. After some algebra it follows for any two items i and j that

$$\omega_{ij}(\theta) = \frac{P_i(\theta)Q_j(\theta)}{Q_i(\theta)P_j(\theta)} = \frac{(1-\gamma e^{b_i})(e^{b_j} - \delta)}{(e^{b_i} - \delta)(1-\gamma e^{b_j})},$$

which is independent of θ. The IRF for item $i, i = 1, \ldots, k$, in Equation 18 is a continuous, strictly increasing four-parameter logistic IRF and, therefore, also nondecreasing in θ. This completes the proof of Theorem 5. □

If under this general parametric CLORs model both $\delta > 0$ and $\gamma > 0$, then each IRF has both a unique lower asymptote unequal to zero and a unique upper asymptote unequal to one. This means that the model is not equivalent to the one-parameter logistic Rasch model. As an example, for a set of nine items with $\delta = 0.06$ and $\gamma = 0.05$, selected b-parameter values and the corresponding c- and g-parameter values are given in Table 1. Graphical representations of the IRFs of the nine items, with parameter values given in Table 1, are shown in Figure 2.

The general parametric CLORs model has two special cases that are generalizations of the one-parameter logistic Rasch model (Hessen, 2004). The first special case with $\delta > 0$ and $\gamma = 0$ is a special case of the three-parameter logistic model (Birnbaum, 1968). The three-parameter logistic IRF for item $i, i = 1, \ldots, k$, is given by the four-parameter logistic IRF, with $g_i = 1$, for item $i, i = 1, \ldots, k$ (Birnbaum, 1968).

Theorem 6. The assumption of LI and a set of k three-parameter logistic IRFs (Birnbaum, 1968) define a special case of the general parametric CLORs model, if and only if $c_i e^{a_i b_i} = \delta > 0$ and $a_i = a = 1$, for all $i, i = 1, \ldots, k$.

TABLE 1.
The b-, c-, and g-parameter values of nine items under the general parametric CLORs model, with $\delta = 0.06$ and $\gamma = 0.05$.

Item	1	2	3	4	5	6	7	8	9
b_i	−2.00	−1.50	−1.00	−0.50	0.00	0.50	1.00	1.50	2.00
c_i	0.44	0.27	0.16	0.10	0.06	0.03	0.02	0.01	0.01
g_i	0.99	0.99	0.98	0.97	0.95	0.92	0.87	0.78	0.63

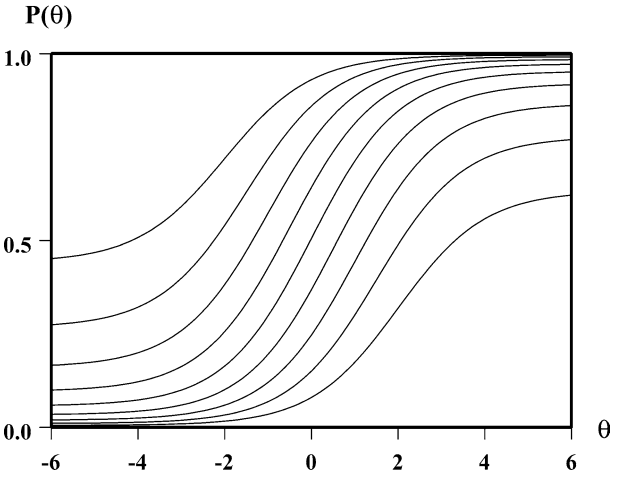

FIGURE 2.
IRFs of nine items under the general parametric CLORs model, with parameter values given in Table 1.

Proof. The ORF for any two items i and j with three-parameter logistic IRFs is

$$\omega_{ij}(\theta) = \frac{(1-c_j)(c_i + e^{a_i(\theta-b_i)})}{(1-c_i)(c_j + e^{a_j(\theta-b_j)})}.$$

This function $\omega_{ij}(\theta)$ is equal to the constant α_{ij} for all θ, if and only if the function

$$\lambda_{ij}(\theta) = \frac{c_i + e^{a_i(\theta-b_i)}}{c_j + e^{a_j(\theta-b_j)}}$$

is equal to $\frac{(1-c_i)}{(1-c_j)}\alpha_{ij}$, for all θ. The function $\lambda_{ij}(\theta)$ is equal to a constant for all θ, if and only if its first derivative is equal to zero for all θ. Therefore, the set of solutions for the item parameters is obtained by solving the identity

$$\lambda'_{ij} = \frac{d}{d\theta}\left[\frac{c_i + e^{a_i(\theta-b_i)}}{c_j + e^{a_j(\theta-b_j)}}\right] \equiv 0$$

for the item parameters. Applying the quotient rule, and setting the result equal to zero, yields

$$\frac{c_i + e^{a_i(\theta-b_i)}}{c_j + e^{a_j(\theta-b_j)}} \equiv \frac{a_i e^{a_i(\theta-b_i)}}{a_j e^{a_j(\theta-b_j)}}.$$

Next, some algebra shows that

$$a_i c_j e^{a_j b_j} \cdot e^{-a_j\theta} + a_i \equiv a_j c_i e^{a_i b_i} \cdot e^{-a_i\theta} + a_j,$$

which gives the set of solutions $\{a_i c_j e^{a_j b_j} = a_j c_i e^{a_i b_i}, -a_j = -a_i, a_i = a_j\}$. Solving the necessary restriction $c_i e^{a_i b_i} = \delta$ for c_i yields $c_i = \delta e^{-a_i b_i}$. Now, applying the restrictions $g_i = 1$, $c_i = \delta e^{-a_i b_i}$, and $a_i = a = 1$, for all i, to the IRF in Equation 17 gives

$$P_i(\theta) = \frac{\delta e^{-b_i} + e^{(\theta-b_i)}}{1 + e^{(\theta-b_i)}} = \frac{\delta + e^{\theta}}{e^{b_i} + e^{\theta}}, \quad \text{for item } i,\ i = 1, \ldots, k, \tag{19}$$

TABLE 2.
The b- and c-parameter values of nine items under the δ-model with $\delta = 0.05$.

Item	1	2	3	4	5	6	7	8	9
b_i	−2.00	−1.50	−1.00	−0.50	0.00	0.50	1.00	1.50	2.00
c_i	0.37	0.22	0.14	0.08	0.05	0.03	0.02	0.01	0.01

which is equivalent to the IRF for item i, $i = 1, \ldots, k$, of the general parametric CLORs model with $\gamma = 0$ (Hessen, 2004). □

The model defined by LI and the IRF in Equation 19 will be called the δ-model. This δ-model is not equivalent to the one-parameter logistic Rasch model, because when $\delta > 0$, each IRF has a unique lower asymptote, and $c_i > 0$, for all i. Note that the c-parameter value is negatively related to the b-parameter value, that is, if $b_i > b_j$, then $c_i < c_j$, for any two items i and j. Moreover, it follows that if $\delta > 0$ and $c_i = c$, for all i, then $c > 0$ and $b_i = \ln \delta - \ln c = b$, for all i. In the latter case, all items have the same IRF. Therefore, the δ-model cannot coincide with one-parameter logistic models with an invariant or fixed guessing parameter (Cressie and Holland, 1983; Weitzman, 1996).

As an example, for a set of nine items under the δ-model with $\delta = 0.05$, selected b-parameter values and the corresponding c-parameter values are given in Table 2. Graphical representations of the IRFs of the nine items under the δ-model with parameter values given in Table 2 are shown in Figure 3.

Another three-parameter special case of the four-parameter logistic model defined by LI and the IRF in Equation 17 can be obtained by restricting all lower limits to zero. The four-parameter logistic IRF with $c_i = 0$, for item i, $i = 1, \ldots, k$, yields

$$P_i(\theta) = \frac{g_i e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}, \quad \text{for item } i, i = 1, \ldots, k. \tag{20}$$

Theorem 7. The assumption of LI and a set of k logistic IRFs given in Equation 20 define a special case of the general parametric CLORs model, if and only if $(1 - g_i)e^{-a_i b_i} = \gamma > 0$ and $a_i = a = 1$, for all i, $i = 1, \ldots, k$.

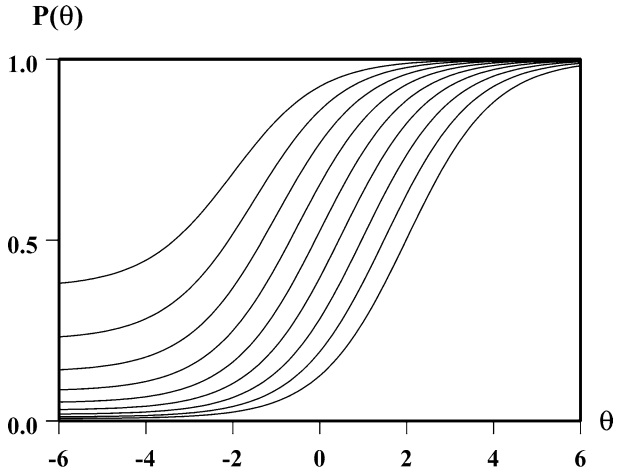

FIGURE 3.
IRFs of nine items under the δ-model with parameter values given in Table 2.

TABLE 3.
The b- and g-parameter values of nine items under the γ-model with $\gamma = 0.05$.

Item	1	2	3	4	5	6	7	8	9
b_i	−2.00	−1.50	−1.00	−0.50	0.00	0.50	1.00	1.50	2.00
g_i	0.99	0.99	0.98	0.97	0.95	0.92	0.86	0.78	0.63

Proof. The necessary restrictions can be obtained in a similar way as for the proof of Theorem 6. Solving the necessary restriction $(1 - g_i)e^{-a_i b_i} = \gamma$ for g_i, yields $g_i = 1 - \gamma e^{a_i b_i}$. Applying the necessary restrictions $g_i = 1 - \gamma e^{a_i b_i}$ and $a_i = a = 1$, for all i, $i = 1, \ldots, k$, to the IRF in Equation 20 gives

$$P_i(\theta) = \frac{(1 - \gamma e^{b_i})e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} = \frac{(1 - \gamma e^{b_i})e^{\theta}}{e^{b_i} + e^{\theta}}, \quad \text{for item } i, i = 1, \ldots, k, \tag{21}$$

which is equivalent to the IRF for item i, $i = 1, \ldots, k$, of the general parametric CLORs model with $\delta = 0$ (Hessen, 2004). □

The model defined by LI and the IRF in Equation 21 will be called the γ-model. This model is also not equivalent to the one-parameter logistic Rasch model, because when $\gamma < e^{-b_i}$, each IRF has a unique upper asymptote, and $g_i < 1$, for all i, $i = 1, \ldots, k$.

The parameter g_i is the upper asymptote of the IRF of item i, $i = 1, \ldots, k$, while the lower asymptote equals zero for all IRFs. This model may be of theoretical interest only, although in practice the inclusion of a g-parameter could account for the possible phenomenon that high-ability examinees do not answer test items correctly when they have information beyond that assumed by the item-constructor.

As an example, for a set of nine items with $\gamma = 0.05$, selected b-parameter values and the corresponding g-parameter values are given in Table 3. Graphical representations of the IRFs of the nine items with item parameter values given in Table 3 are given in Figure 4.

In this section, one nonparametric and three parametric special cases of the general CLORs model that are not equivalent to the one-parameter logistic Rasch model have been discussed. The following theorem relates the one-parameter logistic Rasch model to these CLORs models.

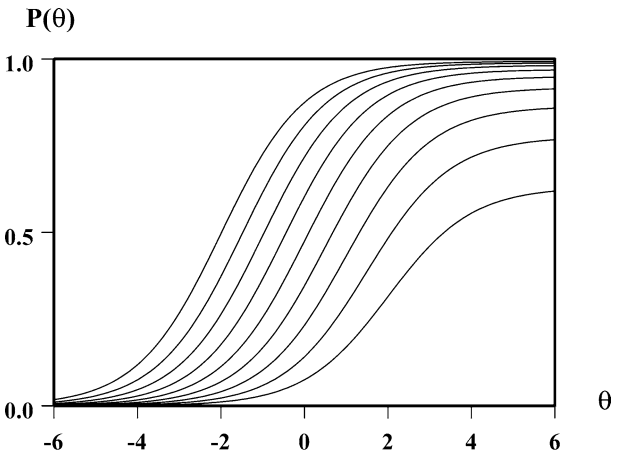

FIGURE 4.
IRFs of nine items under the γ-model with parameter values given in Table 3.

Theorem 8. The one-parameter logistic Rasch model is a special case of: (1) the general CLORs model defined in Subsection 2.1, (2) the nonparametric monotone CLORs model defined in Subsection 2.3, (3) the general parametric CLORs model defined in Theorem 5, (4) the δ-model defined in Theorem 6, and (5) the γ-model defined in Theorem 7.

To prove Theorem 8, it suffices to show that the one-parameter logistic Rasch model is a special case of both the δ-model and the γ-model, because the latter two models have been shown to be special cases of all other CLORs models listed in Theorem 8.

Proof. If $\delta = 0$ in the IRF of the δ-model, then $c_i = 0$, for all i, $i = 1, \ldots, k$, and the resulting model equals the one-parameter logistic Rasch model. Thus, the one-parameter logistic Rasch model is a special case of the δ-model. Furthermore, if $\gamma = 0$ in the IRF of the γ-model, then $g_i = 1$, for all i, $i = 1, \ldots, k$, and again, the resulting model equals the one-parameter logistic Rasch model. So the one-parameter logistic Rasch model is also a special case of the γ-model. □

Since under all CLORs models the total score is a sufficient statistic for the latent trait, under certain conditions the MH procedure can be used to investigate DIF. These conditions are studied in the next section.

3. CLORs and the MH Procedure

In the MH procedure, the null hypothesis is tested whether the manifest between-groups odds-ratios for a studied item j are equal to one at all total score levels. The two groups of examinees are usually called the reference group (majority) and the focal group (minority). If the reference group is denoted by $G = 1$ and the focal group by $G = 2$, then the MH null hypothesis is formulated as

$$H_0 : \frac{\frac{P(X_j=1|X_+=x_+,G=1)}{P(X_j=0|X_+=x_+,G=1)}}{\frac{P(X_j=1|X_+=x_+,G=2)}{P(X_j=0|X_+=x_+,G=2)}} = \beta_j^{MH} = 1, \quad \text{for all } x_+. \tag{22}$$

The MH null hypothesis is tested against the alternative hypothesis of a constant manifest odds-ratio unequal to one at all total score levels; that is,

$$H_1 : \beta_j^{MH} \neq 1, \quad \text{for all } x_+. \tag{23}$$

In this section, the manifest MH odds-ratios conditional on the total score are rewritten in terms of CLORs. The manifest MH odds-ratios in terms of CLORs allow for an explicit look at four important cases that can arise in any DIF analysis.

First, for an arbitrary population of examinees the manifest probabilities of a correct response and that of an incorrect response, both conditional on the total score, are derived from the manifest probability given in Equation 2. By using the assumption of LI in Equation 3 and the result in Equation 10 the manifest probability in Equation 2 can be rewritten as

$$P(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{k-1} \alpha_{ik}^{x_i} \int V_k(\theta)^{x_+} \prod_{i=1}^{k} Q_i(\theta) \, \mathrm{d}F(\theta), \quad \text{for all } \mathbf{x}. \tag{24}$$

Again, let R_{x_+} be the set of all response patterns for which $X_+ = x_+$; that is, $R_{x_+} = \{\mathbf{x}; \sum_{i=1}^{k} x_i = x_+\}$, and let U_{x_+-1} be the set of all response patterns for which $\sum_{i \neq j}^{k} X_i = x_+ - 1$; that is, $U_{x_+-1} =$

$\{\mathbf{x}; \sum_{i\neq j}^{k} x_i = x_+ - 1\}$. Then, the manifest probability of giving a correct response to item j conditional on the total score is

$$\begin{aligned} P(X_j = 1 \mid X_+ = x_+) &= \frac{P(X_j = 1, X_+ = x_+)}{P(X_+ = x_+)} \\ &= \frac{\alpha_{jk} \sum_{U_{x_+-1}} \prod_{i\neq j}^{k-1} \alpha_{ik}^{x_i} \int V_k(\theta)^{x_+} \prod_{i=1}^{k} Q_i(\theta) \mathrm{d}F(\theta)}{\sum_{R_{x_+}} \prod_{i=1}^{k-1} \alpha_{ik}^{x_i} \int V_k(\theta)^{x_+} \prod_{i=1}^{k} Q_i(\theta) \mathrm{d}F(\theta)} \\ &= \frac{\alpha_{jk} \sum_{U_{x_+-1}} \prod_{i\neq j}^{k-1} \alpha_{ik}^{x_i}}{\sum_{R_{x_+}} \prod_{i=1}^{k-1} \alpha_{ik}^{x_i}}, \quad \text{for all } x_+. \end{aligned} \tag{25}$$

Similarly, if U_{x_+} is the set of all response patterns for which $\sum_{i\neq j}^{k} X_i = x_+$; that is, $U_{x_+} = \{\mathbf{x}; \sum_{i\neq j}^{k} x_i = x_+\}$, then the manifest probability of an incorrect response to item j conditional on the total score is

$$P(X_j = 0 \mid X_+ = x_+) = \frac{\sum_{U_{x_+}} \prod_{i\neq j}^{k-1} \alpha_{ik}^{x_i}}{\sum_{R_{x_+}} \prod_{i=1}^{k-1} \alpha_{ik}^{x_i}}, \quad \text{for all } x_+. \tag{26}$$

Next, it follows that the odds for a correct response on item j conditional on the total score is equal to

$$\frac{P(X_j = 1 \mid X_+ = x_+)}{P(X_j = 0 \mid X_+ = x_+)} = \frac{\alpha_{jk} \sum_{U_{x_+-1}} \prod_{i\neq j}^{k-1} \alpha_{ik}^{x_i}}{\sum_{U_{x_+}} \prod_{i\neq j}^{k-1} \alpha_{ik}^{x_i}}, \quad \text{for all } x_+. \tag{27}$$

Finally, dividing the odds of a correct response in the reference group ($G = 1$) by the odds of a correct response in the focal group ($G = 2$) yields the manifest between-groups MH odds-ratios under the general CLORs model; that is,

$$\beta_j^{MH} = \frac{\alpha_{jk1} \left(\sum_{U_{x_+-1}} \prod_{i\neq j}^{k-1} \alpha_{ik1}^{x_i} \right) \left(\sum_{U_{x_+}} \prod_{i\neq j}^{k-1} \alpha_{ik2}^{x_i} \right)}{\alpha_{jk2} \left(\sum_{U_{x_+-1}} \prod_{i\neq j}^{k-1} \alpha_{ik2}^{x_i} \right) \left(\sum_{U_{x_+}} \prod_{i\neq j}^{k-1} \alpha_{ik1}^{x_i} \right)}, \quad \text{for all } x_+, \tag{28}$$

where α_{ik1}, for $i = 1, \ldots, k-1$, are the CLORs in the reference group, and α_{ik2}, for $i = 1, \ldots, k-1$, are the CLORs in the focal group. In what follows, four familiar hypothetical situations or cases are discussed that can take place in practice when the MH procedure is used to test for DIF in the 'studied item' j. From now on, subscript 1 is used to denote the reference group and subscript 2 is used to denote the focal group.

Case 1: All k items are free of DIF. If all items are free of DIF; that is, if

$$P_{i1}(\theta) = P_{i2}(\theta) = P_i(\theta), \quad \text{for all } \theta \text{ and all } i, \tag{29}$$

then all $k-1$ ORFs between the reference item k and any of the remaining items (including the studied item j) are invariant between groups; that is, $\omega_{ik1}(\theta) = \omega_{ik2}(\theta) = \omega_{ik}(\theta)$, for all θ and $i = 1, \ldots, k-1$, because then

$$\frac{P_{i1}(\theta) Q_{k1}(\theta)}{Q_{i1}(\theta) P_{k1}(\theta)} = \frac{P_{i2}(\theta) Q_{k2}(\theta)}{Q_{i2}(\theta) P_{k2}(\theta)} = \frac{P_i(\theta) Q_k(\theta)}{Q_i(\theta) P_k(\theta)}, \quad \text{for all } \theta \text{ and } i = 1, \ldots, k-1. \tag{30}$$

Since under a CLORs model, the ORFs in Equation 30 are all constant and independent of θ, invariance of the IRFs between the two groups, given in Equation 29, implies

$$\alpha_{ik1} = \alpha_{ik2} = \alpha_{ik}, \quad \text{for } i = 1, \ldots, k-1. \tag{31}$$

Substituting both α_{ik1} and α_{ik2} by α_{ik}, for $i = 1, \ldots, k-1$, in Equation 28 yields a manifest MH odds-ratio equal to one at each total score level; that is,

$$\beta_j^{MH} = 1, \quad \text{for all } x_+. \tag{32}$$

This is a 'true-negative' case, in which the manifest MH odds-ratio correctly indicates that the studied item is a non-DIF item. So, when the MH Chi-square test statistic is used to test item j for DIF in a practical situation in which all items, including item j, are free of DIF, then the actual level of significance will be asymptotically equal to the nominal level of significance.

Case 2: All k items are DIF items. Note that $\beta_j^{MH} = 1$, for all x_+, can also be true without invariance of the IRFs between the two groups for any item in the matching criterion, because $\alpha_{ik1} = \alpha_{ik2} = \alpha_{ik}$, for $i = 1, \ldots, k-1$, does not imply that $P_{i1}(\theta) = P_{i2}(\theta)$, for all θ and $i = 1, \ldots, k$. For example, if

$$\frac{P_{i1}(\theta)Q_{i2}(\theta)}{Q_{i1}(\theta)P_{i2}(\theta)} = \alpha \neq 1, \quad \text{for all } \theta \text{ and } i = 1, \ldots, k, \tag{33}$$

then all items are uniform DIF items according to the definition of Hanson (1998), with the same amount of DIF for all items, expressed by α. From this it follows that

$$\frac{P_{i1}(\theta)Q_{i2}(\theta)}{Q_{i1}(\theta)P_{i2}(\theta)} = \frac{P_{k1}(\theta)Q_{k2}(\theta)}{Q_{k1}(\theta)P_{k2}(\theta)}, \quad \text{for } i = 1, \ldots, k-1, \tag{34}$$

which is equivalent to

$$\frac{P_{i1}(\theta)Q_{k1}(\theta)}{Q_{i1}(\theta)P_{k1}(\theta)} = \frac{P_{i2}(\theta)Q_{k2}(\theta)}{Q_{i2}(\theta)P_{k2}(\theta)}, \quad \text{for } i = 1, \ldots, k-1, \tag{35}$$

and under a CLORs model this means that

$$\alpha_{ik1} = \alpha_{ik2}, \quad \text{for } i = 1, \ldots, k-1. \tag{36}$$

Consequently, $\beta_j^{MH} = 1$, for all x_+, in Equation 28. This is a 'false-negative' case in which the manifest MH odds-ratio incorrectly indicates that the studied item is free of DIF. When these conditions are true in a practical DIF testing situation, the MH procedure is incapable of detecting DIF, which is actually present for all items. This situation will be explained in more detail in the worst case scenario example at the end of this section.

Case 3: Only the studied item is free of DIF. If the CLOR between the studied item j and the arbitrary reference item k does not have the same value in the two groups, and if all other CLORs are invariant between the groups; that is,

$$\alpha_{jk1} \neq \alpha_{jk2}, \text{ and } \alpha_{ik1} = \alpha_{ik2}, \quad \text{for all } i \neq j, \tag{37}$$

which implies that $\alpha_{ji1} \neq \alpha_{ji2}$, for all $i \neq j$; then under the general CLORs model, the manifest MH odds-ratio is

$$\beta_j^{MH} = \frac{\alpha_{jk1}}{\alpha_{jk2}} = \frac{\frac{V_{j1}(\theta)}{V_{k1}(\theta)}}{\frac{V_{j2}(\theta)}{V_{k2}(\theta)}}, \quad \text{for all } x_+. \tag{38}$$

If in this case the studied item j is free of DIF, but the reference item k is not free of DIF (which in this case means that only the studied item j is free of DIF), then $V_{j1}(\theta) \equiv V_{j2}(\theta)$ and $V_{k1}(\theta) \neq V_{k2}(\theta)$. Consequently, the manifest MH odds-ratio of the studied non-DIF item j is equal to

$$\beta_j^{MH} = \frac{\alpha_{jk1}}{\alpha_{jk2}} = \frac{\frac{V_{j1}(\theta)}{V_{k1}(\theta)}}{\frac{V_{j2}(\theta)}{V_{k2}(\theta)}} = \frac{V_{k2}(\theta)}{V_{k1}(\theta)} = \frac{P_{k2}(\theta)Q_{k1}(\theta)}{Q_{k2}(\theta)P_{k1}(\theta)} = \frac{1}{\alpha_k} \neq 1, \text{ for all } x_+, \tag{39}$$

which does not reflect a constant amount of DIF in item j, but the common amount of DIF actually present for each of the remaining items in the total score, which are represented by the reference item k. This is a 'false-positive' case in which the manifest MH odds-ratio incorrectly indicates that DIF is present in a non-DIF item.

Case 4: Only the studied item is a DIF item. In the situation in which the reference item k is free of DIF and the studied item j is a DIF item, we have $V_{k1}(\theta) \equiv V_{k2}(\theta)$ and $V_{j1}(\theta) \neq V_{j2}(\theta)$. If at the same time all CLORs are invariant between groups except the CLOR between the studied item j and the reference item k, then except for the studied item j all other items in the total score are free of DIF. In this situation, it follows that the manifest MH odds-ratio is equal to

$$\beta_j^{MH} = \frac{\alpha_{jk1}}{\alpha_{jk2}} = \frac{\frac{V_{j1}(\theta)}{V_{k1}(\theta)}}{\frac{V_{j2}(\theta)}{V_{k2}(\theta)}} = \frac{V_{j1}(\theta)}{V_{j2}(\theta)} = \frac{P_{j1}(\theta)Q_{j2}(\theta)}{Q_{j1}(\theta)P_{j2}(\theta)} = \alpha_j \neq 1, \tag{40}$$

at each total score level. This is a 'true-positive' case in which the manifest MH odds-ratio correctly indicates that DIF is present in an item that is really a DIF item.

If the conditions of Case 4 hold under the CLORs model, then it follows that the studied item j can only be a uniform DIF item, which corresponds to a shift of the difficulty parameter under the Rasch model (Fischer, 1993). Since α_{jk1} and α_{jk2} are constants and $\alpha_{jk1} \neq \alpha_{jk2}$, α_j is also a constant unequal to one and independent of θ, which is the definition of uniform DIF (Hanson, 1998). In constrast to the Rasch model, however, under the general CLORs model uniform DIF does not imply parallel DIF, defined as $P_{i1}(\theta) = P_{i2}(\theta + \epsilon)$ for all θ and $\epsilon \neq 0$. Since under the latter model lower and/or upper asymptotes can differ between the IRFs of the two groups, these IRFs cannot always be horizontal translations of each other.

On the basis of the foregoing results, it can be concluded that only when the IRFs of the reference item k in the two groups are equal and the CLORs between item k and all the other items except for the studied item j are equal in the two groups, then the manifest MH odds-ratios can be used to test the IRT null hypothesis of no-DIF for the studied item j.

Of the four cases discussed, the so-called 'false-positive' and 'false-negative' cases are the undesirable situations. Since in practice the 'false-positive' case can lead to the 'false-negative' one through the deletion of the non-DIF item that is flagged as a DIF item, the 'false-negative' case can be seen as the worst case scenario of the four situations discussed. The 'false-negative' case is especially adverse, because in this case it will be tempting, after the items are all investigated for DIF, to draw the conclusion that none of the items are DIF items while in fact they all are.

Situations under the CLORs model in which two or more CLORs are different for the two groups are more complex and difficult to discuss.

3.1. Worst Case Scenario Example

In the worst case scenario, DIF exists for all items but the CLORs between all item-pairs in the two groups are equal. Consequently, in this situation the MH null hypothesis is satisfied, which is illustrated in the following numerical example.

Suppose a test consisting of three dichotomous items, and that, for computational convenience, within each of the two groups under study the Rasch model holds, which is a special case of the general CLORs model. In both groups, the common discrimination parameter is equal to 1. Item 3 is arbitrarily taken as the reference item. In both groups, the CLOR between item 1 and item 3 is equal to $\alpha_{13} = e^{b_3 - b_1} = 3$, and between item 2 and item 3, $\alpha_{23} = e^{b_3 - b_2} = 5$. So, for each item the value of the manifest MH odds-ratio is equal to $\beta_k^{MH} = 1$ at each total score level. If in group 1 the item difficulty parameter value of the third item is $b_3 = 2$, then in this group $b_1 = 2 - \ln 3 = 0.901$ and $b_2 = 2 - \ln 5 = 0.391$. When in group 2 (denoted by an asterisk) the difficulty parameter value for the third item is $b_3^* = 1$, then in this group $b_1^* = 1 - \ln 3 = -0.099$ and $b_2^* = 1 - \ln 5 = -0.609$. This means that for each item the signed area (SA_i) (Raju, 1988) has the value 1 ($SA_3 = b_3 - b_3^* = 2 - 1 = 1$; $SA_2 = b_2 - b_2^* = 0.391 - (-0.609) = 1$; and $SA_1 = b_1 - b_1^* = 0.901 - (-0.099) = 1$). This value indicates a large difference between the corresponding difficulty parameter values in the two groups for each item. So, for each item in this example the MH null hypothesis is satisfied while each item has a large amount of DIF.

4. Discussion

In the present paper, it has been shown that given the assumption of LI, the condition of CLORs for all pairs of items is equivalent to sufficiency of the total score for the latent trait. Rosenbaum (1984) proposed tests of LI that can be applied without first specifying a parametric form of the IRFs. Rosenbaum (1987) also showed that the simple assumption of a CLOR for a single pair of items, in his terminology called proportional latent odds, places testable restrictions on the observable distribution of the item scores. The Chi-square test of independence he proposed for checking whether a pair of items has a CLOR is easy to implement, and probably works well with samples of moderate size.

In this paper also two simple observable consequences, which can be used to evaluate whether the general CLORs model can represent the data structure, are derived. WII is one of these observable consequences and is related to IIO. Manifest monotonicity (Junker and Sijtsma, 2000) is the second observable consequence. Without additionally assuming latent monotonicity, LI of the item scores and CLORs for all item-pairs are shown to be sufficient conditions for manifest monotonicity or WSI (Scheiblechner, 1995). This means that manifest monotonicity does not imply latent monotonicity.

Although both the nonparametric and parametric generalizations of the Rasch model still seem very restrictive, they will be more flexible to empirical data than the Rasch model. Both parameter estimation and goodness-of-fit procedures must be developed to determine whether these models fit significantly better to empirical data than the Rasch model.

An important requirement with which IRT models must comply in general is that they give a realistic description of item response behavior of examinees. In the models with lower asymptotes, the relation between the c-parameter and the location parameter does not seem to be unrealistic for cognitive tests when they are interpreted as a guessing parameter and a difficulty parameter, respectively. It is plausible that the probability of guessing the right answer is lower when an item is more difficult. For example, in case of multiple-choice items with one correct answer and a fixed number of incorrect answer categories, the guessing parameter values should in theory be equal for all items. However, in practice, these values are often not equal to each other. An explanation for this is the varying attractiveness of the incorrect answer categories between items. When an item has attractive incorrect answer categories that are almost correct, then the item will be more difficult and the probability of guessing the right answer will be smaller than in the case of incorrect answer categories that are obviously wrong. For other tests, such as dichotomously scored personality questionnaires or other psychological tests, for which the positively keyed answer is indicative for more of a certain opinion, attitude, emotional state, and so on, the lower

asymptote cannot be interpreted as a guessing probability. However, there is no a priori reason why the IRFs of such items should have lower asymptotes equal to zero and upper asymptotes equal to one.

For DIF detection, the sufficiency of the total score for the latent trait means that the total score captures all relevant information in the latent trait for the studied item, and that the total score can be used as a matching variable in an observed score-based DIF detection procedure (Millsap and Meredith, 1992). So, if the data follow the general CLORs model or one of its special cases, including the Rasch model, then observed score-based DIF detection methods in general and the MH procedure in particular can be diagnostic of DIF. In practice, these methods should still be applied with caution, because of the 'false-negative' and 'false-positive' cases discussed in one of the previous sections. The worst case scenario deserves special attention. When under one of the CLORs models all items are DIF items with approximately the same amount of uniform DIF in favor of the reference group while the reference and the focal groups have the same true mean latent ability value, then the total score mean in the reference group will be higher than the total score mean in the focal group. So in that situation, there are no true between-group differences, while it might be concluded on the basis of the observed scores that there are. The groups only differ with respect to something else that is also systematically measured by the test in addition to the intended latent trait. This is a familiar validity problem which cannot be solved by statistical methods. Therefore, the contents of the items should always be thoroughly investigated.

References

Barton, M.A., & Lord, F.M. (1981). An upper asymptote for the three-parameter logistic item-response model. *Research Bulletin 81-20*. Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Cressie, N., & Holland, P.W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika, 48,* 129–141.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests.* Bern: Huber. (Introduction to the theory of psychological tests).

Fischer, G.H. (1993). Notes on the Mantel–Haenszel procedure and another chi-squared test for the assessment of DIF. *Methodika, 7,* 88–100.

Fischer, G.H. (1995a). Derivations of the Rasch model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments and Applications* (pp. 15–38). Berlin Heidelberg, New York: Springer.

Fischer, G.H. (1995b). Some neglected problems in IRT. *Psychometrika, 60,* 459–487.

Grayson, D.A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53,* 383–392.

Hanson, B.A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23,* 244–253.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., & Junker, B.W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62,* 331–347.

Hessen, D.J. (2004). A new class of parametric IRT models for dichotomous item scores. *Journal of Applied Measurement, 5*, 385–397.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika, 59,* 77–79.

Junker, B.W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21,* 1359–1378.

Junker, B.W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24,* 65–81.

Lehmann, E.L. (1994). *Testing statistical hypotheses.* New York: Wiley/Chapman & Hall.

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

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

Meredith, W., & Millsap, R.E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57,* 289–311.

Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Millsap, R.E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16,* 389–402.
Mokken, R.J. (1971). *A theory and procedure of scale analysis.* Berlin: De Gruyter.
Mokken, R.J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417–430.
Mokken, R.J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion". *Applied Psychological Measurement, 10,* 279–285.
Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495–502.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen and Lydiche.
Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–435.
Rosenbaum, P.R. (1987). Comparing item characteristic curves. *Psychometrika, 52,* 217–233.
Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika, 60,* 281–304.
Sijtsma, K., & Junker, B.W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49,* 79–105.
Weitzman, R.A. (1996). The Rasch model plus guessing. *Educational and Psychological Measurement, 56,* 779–790.
Zwick, R. (1990). When do item response function and Mantel–Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–197.

Manuscript received 14 NOV 2002
Final version received 27 JAN 2004