Measurement and Research Department Reports 91-2

Testing Rasch Models For Polytomous Items: With An Example Concerning Detection Of Item Bias

C.A.W. Glas

Cito

Testing Rasch Models For Polytomous Items: With An Example Concerning Detection Of Item Bias.

C.A.W. Glas

Cito
Arnhem, 1991

abstract

Recently Glas and Verhelst (1989) introduced a general theoretical framework for the construction of asymptotically χ^2-distributed test statistics for item response models. As an example they proposed two statistics for the partial credit model with a normal ability distribution. Also the statistics for the Rasch model for dichotomous items proposed by Glas (1988) fit into the given framework.
In the present paper the theory is applied to the construction of a statistic for some item response models for polytomous items, including the partial credit model. The statistic is defined in a conditional maximum likelihood (CML) framework.
An example concerning detection of item bias will be included.

Key words: item response model, Rasch model, conditional maximum likelihood, model test, item bias.

1. Rasch models for polytomous items.

Consider the response of a person, indexed n, to an item, indexed i, which has m(i)+1 response categories indexed j = 0,1,...,m(i). The response of person n wil be represented by an m(i)-dimensional vector of stochastic variables $\boldsymbol{X}_{ni}$ with elements

$$X_{nij} = \begin{cases} 1 \text{ if person n scores in category j on item i,} \\ 0 \text{ if this is not the case,} \end{cases}$$

for j = 1,...,m(i) . So if the respondent scores in category j = 0, $\boldsymbol{X}_{ni}=\mathbf{0}$.

Andersen (1973a) has shown that if there exists a (vector-valued) minimal sufficient statistic $R_n(\boldsymbol{X}_{n1},...,\boldsymbol{X}_{ni},...,\boldsymbol{X}_{nk})$ for the (vector-valued) person parameter $\boldsymbol{\vartheta}_n$, and the sufficient statistic is symmetric in its arguments, the multidimensional Rasch model (Rasch, 1961) necessarily follows. In the multidimensional Rasch model it is assumed that, for all items, response categories with the same index are associated with the same response tendency, i.e., with the same element of $\boldsymbol{\vartheta}_n$.

Further it is assumed that every item relates to the same set of response tendencies and m(i)= m for i = 1,...,k. Let $\boldsymbol{\eta}_i' \stackrel{d}{=} (\eta_{i1},...,\eta_{ij},...,\eta_{im})$, where η_{ij} is a parameter of item i associated with category j, and $\boldsymbol{\vartheta}_n' \stackrel{d}{=} (\vartheta_{n1},...,\vartheta_{nj},...,\vartheta_{nm})$, where ϑ_{nj} is a parameter of person n associated with category j. The main result derived by Andersen can be summarized by saying that if, for j = 1,...,m , $\sum_i X_{nij}$ is sufficient for ϑ_{nj}, the model must have the form

$$\Pr(\boldsymbol{X}_{ni} = \mathbf{x}_{ni} \mid \boldsymbol{\vartheta}_n, \boldsymbol{\eta}_i) = \frac{\exp\left(\sum_{j=1}^{m} x_{nij}(\vartheta_{nj} - \eta_{ij})\right)}{1 + \sum_{h=1}^{m} \exp(\vartheta_{nh} - \eta_{ih})} \quad . \tag{1}$$

In the sequel the response categories j = 0,1,2,...etc. will be associated with the item scores 0,1,2,...etc.

Andersen (1977) has studied a unidimensional version of the model given by (1) and has shown that if $R_n \stackrel{d}{=} \sum_{i,j} j X_{nij}$ is a sufficient statistic for ϑ_n that is symmetric in its

arguments, the model is necessarily given by

$$\Pr(\boldsymbol{X}_{ni} = x_{ni} \mid \vartheta_n, \eta_i) = \frac{\exp(\sum_{j=1}^{m} x_{nij}(j\vartheta_n - \eta_{ij}))}{1 + \sum_{h=1}^{m} \exp(h\vartheta_n - \eta_{ih})} \quad . \tag{2}$$

Further Andersen shows that the so-called "equidistant scoring rule" is the only scoring rule that allows for a minimal sufficient statistic for the person parameter, which is symmetric in its arguments. The equidistant scoring rule prescribes that the difference between two item scores associated with two response categories is constant for all categories. For example, if $m = 3$, the category weights $\{0,1,2,3\}$ or $\{0,2,4,6\}$ are compatible with equidistant scoring, while the category weights $\{0,2,3,5\}$ are not.

A different derivation of the unidimensional Rasch model for polytomous items is given by Masters (1982). This version of the model, called the partial credit model, is derived from the assumption that every category j $(j > 0)$ of an item can be seen as a step which is either taken, or not taken by the respondent. It is assumed that the probability of a person scoring in category j rather than scoring in category $j-1$ is a logistic function of a person parameter ϑ_n and a parameter δ_{ij} associated with category j of item i. Thus, if $j > 0$,

$$\Pr(X_{nij} = 1 \mid X_{nij} = 1 \text{ or } X_{ni(j-1)} = 1, \vartheta_n, \delta_{ij}) = \frac{\exp(\vartheta_n - \delta_{ij})}{1 + \exp(\vartheta_n - \delta_{ij})} \quad . \tag{3}$$

Masters (1982) shows that from (3) it follows that the probability of a person with parameter ϑ_n scoring in category j, $j = 1,\ldots,m(i)$, on an item with parameter $\boldsymbol{\delta}_i$, $\boldsymbol{\delta}_i \stackrel{d}{=} (\delta_{i1},\ldots,\delta_{ij},\ldots,\delta_{im(i)})$ is given by

$$\Pr(\boldsymbol{X}_{ni} = x_{ni} \mid \vartheta_n, \boldsymbol{\delta}_i) = \frac{\exp(\sum_{j=1}^{m(i)} x_{nij} (\sum_{p=1}^{j} (\vartheta_n - \delta_{ip})))}{1 + \sum_{h=1}^{m(i)} \exp(\sum_{p=1}^{h} (\vartheta_n - \delta_{ip}))} \quad . \tag{4}$$

Notice that if $m(i) = m$ for $i = 1,\ldots,k$ and the reparameterization $\eta_{ij} = \sum_{p=1}^{j} \delta_{ip}$ is applied, the models defined by (2) and (4) are equivalent, that is, (2) and (4) are alternative definitions of the same model. Although Andersen (1977) derives the model under the assumption that the numbers of response categories of the items are the same,

it will be shown that if this assumption is broadened to include items with different numbers of response categories, minimal sufficient statistics for the parameters will also exist. However, these minimal sufficient statistics are no longer symmetric in their arguments. Generalizing Andersen's results to a broader class of response formats, however, is beyond the scope of the present paper.

One of the main motivations for studying Masters' parameterization of the model is an interpretation of the parameters which is not possible for Andersen's version. This interpretation can be derived from the model for dichotomous items, where the item parameter can be viewed as the point on the latent scale at which the probability of a correct response and the probability of an incorrect response are equal.
An analogous interpretation can also be applied to the model for polytomous items. This is shown as follows. For every item a set of m(i)+1 item characteristic functions are defined by

$$\psi_{ij}(\vartheta_n) \stackrel{d}{=} \Pr(X_{nij}=1 \mid \vartheta_n, \delta_i) = \frac{\exp(\sum_{h=1}^{j}(\vartheta_n - \delta_{ih}))}{1 + \sum_{h=1}^{m(i)}\exp(\sum_{p=1}^{h}(\vartheta_n - \delta_{ip}))} , \tag{5}$$

for $j=1,\ldots,m(i)$ and

$$\psi_{i0}(\vartheta_n) \stackrel{d}{=} \Pr(X_{ni}=0 \mid \vartheta_n, \delta_i) = \frac{1}{1 + \sum_{h=1}^{m(i)}\exp(\sum_{p=1}^{h}(\vartheta_n - \delta_{ip}))} . \tag{6}$$

It can be easily verified that, for $j=1,\ldots,m(i)$,

$$\psi_{i(j-1)}(\vartheta) = \psi_{ij}(\vartheta) \Leftrightarrow \vartheta = \delta_{ij} . \tag{7}$$

So δ_{ij} is the boundary value at which the probabilities of scoring in category j and category j-1 are equal.

The notion of defining a dichotomous Rasch model for the probability of scoring in some category j rather than in j-1, and thus defining the associated item parameter as a boundery between two adjacent categories, has been identified by Masters and Wright (1984) as a central theme that unifies a general class of IRT models. As a general formulation they introduce the partial credit model and show that the following models can be written as special cases:

(1) the rating scale model (Andrich, 1978a), i.e., the special case where $\delta_{ij} = \beta_i + \tau_j$,

(2) the binomial trials model (Andrich, 1978b), i.e., the special case where $\delta_{ij} = \beta_i + \ln(j/(m(i)-j+1))$,

(3) the Poisson counts model (Rasch, 1977), i.e., the special case where $\delta_{ij} = \beta_i + \ln(j)$.

Although the models by Andrich (1978a&b) are beyond the main theme of this paper, it will be shown that the similarity between these models and the partial credit model makes it possible to adapt the estimation and testing procedures developed for the latter for use with the former. The Poisson counts model, however, is excluded from this adaption, because it differs from the others in the sense that m(i) is not bounded. The consequences of this feature will become clear later.

2. Another look at conditional maximum likelihood estimation.

Consider a test of k items and let the stochastic vector $\boldsymbol{X}$ represent a response pattern, that is $\boldsymbol{X}' = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_i', \ldots, \boldsymbol{X}_k')$, where $\boldsymbol{X}_i$ stands for a response to item i (if an arbitrary person is considered, the index n will be dropped for convenience and $\boldsymbol{X}_{ni}$ is written as $\boldsymbol{X}_i$). The probability of observing response pattern $\mathbf{x}$ as a function of ϑ and $\boldsymbol{\eta}' = (\eta_1', \ldots, \eta_i', \ldots, \eta_k')$ is given by

$$\Pr(\boldsymbol{X} = \mathbf{x} \mid \vartheta, \boldsymbol{\eta}) = \exp(-\mathbf{x}'\boldsymbol{\eta}) \exp(r(\mathbf{x})\vartheta)\, p_0(\vartheta) \quad , \tag{8}$$

with

$$p_0(\vartheta) \stackrel{d}{=} \prod_{i=1}^{k} \left(1 + \sum_{h=1}^{m(i)} \exp(h\vartheta - \eta_{ih}) \right)^{-1} \quad , \tag{9}$$

and $r(\mathbf{x})$ the sum score associated with response pattern $\mathbf{x}$, that is, $r(\mathbf{x}) \stackrel{d}{=} \sum_i j x_{ij}$. Conditioning on $R(\boldsymbol{X}) \stackrel{d}{=} \sum_i j X_{ij}$ results in the conditional probability

$$\pi_{\mathbf{x}|r} = \Pr(\boldsymbol{X} = \mathbf{x} \mid R(\boldsymbol{X}) = r, \boldsymbol{\eta}) = \frac{\Pr(\boldsymbol{X} = \mathbf{x} \mid \vartheta, \boldsymbol{\eta})}{\sum_{\{\mathbf{x} \mid r(\mathbf{x}) = r\}} \Pr(\boldsymbol{X} = \mathbf{x} \mid \vartheta, \boldsymbol{\eta})}$$

$$= \frac{\exp(-\mathbf{x}'\boldsymbol{\eta})}{\sum_{\{\mathbf{x} \mid r(\mathbf{x}) = r\}} \exp(-\mathbf{x}'\boldsymbol{\eta})} \quad , \tag{10}$$

where $\{\mathbf{x} \mid r(\mathbf{x}) = r\}$ stands for the set of all response patterns leading to sum score r. It will prove convenient to introduce the concept of an elementary function. Let K be the maximum score that can be obtained on the test, so $K = \sum_i m(i)$, and let $\varepsilon_{ij} \stackrel{d}{=} \exp(-\eta_{ij})$. For $r = 0, \ldots, K$, the elementary function of order r is defined by

$$\Gamma_r \stackrel{d}{=} \sum_{\{\mathbf{x} \mid r(\mathbf{x}) = r\}} \prod_{i,j} \varepsilon_{ij}^{x_{ij}} = \sum_{\{\mathbf{x} \mid r(\mathbf{x}) = r\}} \exp(-\mathbf{x}'\boldsymbol{\eta}) \quad . \tag{11}$$

It is assumed that elementary functions of an order less then zero are equal to zero. The following example may clarify this definition. For a test of three items with $m(i) = 2$, for $i = 1, \ldots, 3$ the elementary functions are given by:

$$\Gamma_0 = 1,$$
$$\Gamma_1 = \varepsilon_{11} + \varepsilon_{21} + \varepsilon_{31},$$
$$\Gamma_2 = \varepsilon_{11}\varepsilon_{21} + \varepsilon_{11}\varepsilon_{31} + \varepsilon_{21}\varepsilon_{31} + \varepsilon_{12} + \varepsilon_{22} + \varepsilon_{32},$$
$$\Gamma_3 = \varepsilon_{11}\varepsilon_{21}\varepsilon_{31} + \varepsilon_{12}\varepsilon_{21} + \varepsilon_{11}\varepsilon_{22} + \varepsilon_{12}\varepsilon_{31} + \varepsilon_{11}\varepsilon_{32} + \varepsilon_{22}\varepsilon_{31} + \varepsilon_{21}\varepsilon_{32},$$
$$\Gamma_4 = \varepsilon_{12}\varepsilon_{21}\varepsilon_{31} + \varepsilon_{12}\varepsilon_{22} + \varepsilon_{12}\varepsilon_{32} + \varepsilon_{11}\varepsilon_{22}\varepsilon_{31} + \varepsilon_{11}\varepsilon_{21}\varepsilon_{32} + \varepsilon_{22}\varepsilon_{32},$$
$$\Gamma_5 = \varepsilon_{12}\varepsilon_{22}\varepsilon_{31} + \varepsilon_{12}\varepsilon_{21}\varepsilon_{32} + \varepsilon_{11}\varepsilon_{22}\varepsilon_{32},$$
$$\Gamma_6 = \varepsilon_{12}\varepsilon_{22}\varepsilon_{32}.$$

The computation of elementary functions defined by (11) has been described by Andersen (1972) and Fischer (1974).

Using these definitions, (10) can also be written as

$$\pi_{x|r} = \frac{\prod_{i,j} \varepsilon_{ij}^{x_{ij}}}{\Gamma_r} \quad . \tag{12}$$

From (10) or (12) it can be easily verified that within every score level r, the probabilities $\pi_{x|r}$ sum to one, i.e.,

$$\sum_{\{x \mid r(x) = r\}} \pi_{x|r} = 1 \quad , \tag{13}$$

for $r = 0,\ldots,K$. Further, every respondent displays only one response pattern, and so, conditional on the sum scores, the sampling model is product-multinomial.
Birch (1963) and Haberman (1974) have shown that ML estimation procedures and statistical testing procedures for parametric product-multinomial models can easily be transformed into equivalent procedures for multinomial models. Applied to the present problem, the transformation can be carried out as follows. For $r = 0,\ldots,K$, let N_r be the number of persons in the sample obtaining sum score r. Assume that N_r, for $r = 0,\ldots,K$, has a multinomial distribution defined by the total sample size N and the probabilities $\upsilon_0,\ldots,\upsilon_r,\ldots,\upsilon_K$. Notice that the ML estimate of υ_r is given by $\hat{\upsilon}_r = n_r/N$.
The probability of response pattern x can now be given as $\pi_x \stackrel{d}{=} \upsilon_r \pi_{x|r}$, or

$$\pi_{\mathbf{x}} \stackrel{d}{=} \Pr(\boldsymbol{X}=\mathbf{x} \mid \boldsymbol{\varepsilon}, \upsilon) = \frac{\upsilon_r \prod_{i,j} \varepsilon_{ij}^{x_{ij}}}{\Gamma_r} \quad , \tag{14}$$

with $\boldsymbol{\varepsilon}' \stackrel{d}{=} (\varepsilon_{11},\ldots,\varepsilon_{km(k)})$ and $\boldsymbol{\upsilon}' \stackrel{d}{=} (\upsilon_0,\ldots,\upsilon_r,\ldots,\upsilon_K)$.

Let $\{\mathbf{x}\}$ stand for the set of all possible response patterns on the test. Then the data can be represented by a vector of frequency counts $\boldsymbol{N}$, which has elements $N_{\mathbf{x}}$ for all $\mathbf{x} \in \{\mathbf{x}\}$, where $N_{\mathbf{x}}$ is the number of respondents producing response pattern $\mathbf{x}$. Suppose that the number of possible response patterns, that is, the number of elements in $\{\mathbf{x}\}$, is equal to v. Further $\boldsymbol{\pi}$ is defined as a v-dimensional vector with elements $\pi_{\mathbf{x}}$, for all $\mathbf{x} \in \{\mathbf{x}\}$. Since the probabilities $\pi_{\mathbf{x}}$ sum to one, the vector of frequency counts $\boldsymbol{N}$ has a multinomial distribution defined by N and $\boldsymbol{\pi}$.
With these definitions, the CML estimation procedure can be brought within the well-established framework of parametric multinomial models. The multinomial form of the distribution of $\boldsymbol{N}$ is not only practical for the derivation of estimation equations and asymptotic confidence intervals, it will prove to be essential for the derivation of the distribution of statistics for the evaluation of model fit. It is well-known (see for instance Andersen, 1980, or Barndorff-Nielsen, 1978) that if the distribution function of the data belongs to an exponential family, ML estimation boils down to equating the realizations of the sufficient statistics with their expected values. It will now be shown that this convenient method for deriving estimation equations can also be applied to the Rasch model.

If the distribution function of the counts of the response patterns belongs to an exponential family and the model is parameterized by an s-dimensional vector of parameters $\boldsymbol{\xi}$, the probability of observing $\mathbf{x}$, must have the form

$$\pi_{\mathbf{x}} = b(\mathbf{x})\exp(\boldsymbol{\xi}'\mathbf{t}(\mathbf{x}))/a(\boldsymbol{\xi}) \quad , \tag{15}$$

where $\mathbf{t}(\mathbf{x})$ is a s-dimensional vector which is a function of $\mathbf{x}$ only, $b(\mathbf{x})$ is a function of $\mathbf{x}$ only, and $a(\boldsymbol{\xi})$ is a function of $\boldsymbol{\xi}$ only. If the parameters are linearly independent, that is, if there does not exist a linear transformation of $\boldsymbol{\xi}$ that leaves the probabilities

unchanged, the elements of ξ are called "canonical" or "natural" parameters (see, for instance, Andersen, 1980, p.20). Let T be an $s \times v$ matrix with columns $t(x)$ and let D_π be a diagonal matrix of the elements of π. Both for the derivation of estimation equations and the asymptotic distribution of test statistics, the following lemma will prove convenient.

Lemma 1. $\partial\pi/\partial\xi' = D_\pi T' - \pi\pi'T'$.

Proof. Since all probabilities π_x sum to one, the factor $a(\xi)$ in (15) can be written as $\sum_{\{x\}} b(x)\exp(\xi' t(x))$. Let y be some response vector belonging to $\{x\}$. Then $\partial\pi_y/\partial\xi_j = t_j(y)\pi_y - \pi_y \sum_{\{x\}} t_j(x)\pi_x$, for $j = 1,\ldots,s$ and the result follows. □

In the present section, the lemma is used for deriving estimation equations, the relevance with respect to the derivation of the asymptotic distribution of test statistics will become clear in the next section. Let n be a realization of N. The log-likelihood function of ξ, $\ln L(\xi|n)$, can be written as $\sum_{\{x\}} n_x \ln\pi_x + c$, where c is a constant which does not depend on ξ. The partial derivatives of $\ln L(\xi|n)$ with respect to ξ are given by

$$\partial \ln L(\xi|n)/\partial\xi' = n'D_\pi^{-1}(\partial\pi/\partial\xi') = n'D_\pi^{-1}\left(D_\pi T' - \pi\pi'T'\right) = n'T' - n'D_\pi^{-1}\pi\pi'T' .$$

But $D_\pi^{-1}\pi = 1_v$, with 1_v a v-dimensional vector with all elements equal to one, and, as a result, the estimation equations are given by $n'T' = N\pi'T'$, or more conventionally, $Tn = NT\pi$. However, Tn is an s-dimensional vector of observed sufficient statistics and $NT\pi$ is its expectation, so the celebrated result that in an exponential family ML estimation is equivalent with matching expected and observed values of sufficient statistics has been derived again.

Returning to the Rasch model, for $r = 0,\ldots,K$, let $\omega_r \stackrel{d}{=} \ln(\upsilon_r/\Gamma_r)$ and $\xi' \stackrel{d}{=} (-\eta_{11},\ldots,-\eta_{k,m(k)-1},\omega_0,\ldots,\omega_{K-1})$. Since only K-1 free item parameters can be estimated, $\eta_{km(k)}$ is not included in ξ. For the same reason ω_K is not included in ξ. Notice that in the present case the dimension of ξ, which was defined as s, is equal to 2K-1. Let $t(x)$ be defined by $t(x)' \stackrel{d}{=} (x_1,\ldots,x_{K-1},e'_{r(x)})$, where $r(x)$ is the sum score associated with x and $e'_{r(x)}$ is a K-dimensional vector with all elements equal to zero, except the $(r(x)+1)$-th element, which is equal to one. With these definitions of ξ and $t(x)$, $\pi_x = \exp(t(x)'\xi)$ and it directly follows that the multinomial model defined by N and π_x, for all $x \in \{x\}$, belongs to an exponential family.

To derive the CML estimation equations for the item parameters, the structure of the matrix T must be specified in detail. Let v(r) be the number of response patterns leading to sum score r and let the $(2K+1) \times v$ matrix T^* be defined by

$$T^* = \begin{bmatrix} 0 & X_1 & X_2 & \dots & X_r & \dots & X_{K-1} & x_K \\ 1 & 0' & 0' & \dots & 0' & \dots & 0' & 0 \\ 0 & 1'_{v(1)} & 0' & \dots & 0' & \dots & 0' & 0 \\ 0 & 0' & 1'_{v(2)} & \dots & 0' & \dots & 0' & 0 \\ 0 & 0' & 0' & \dots & 1'_{v(r)} & \dots & 0' & 0 \\ 0 & 0' & 0' & \dots & 0' & \dots & 1'_{v(K-1)} & 0 \\ 0 & 0' & 0' & \dots & 0' & \dots & 0' & 1 \end{bmatrix}, \tag{16}$$

where X_r is a $K \times v(r)$ matrix with as columns all response patterns leading to a sum score r, x_k stands for the response pattern leading to a perfect score and $1_{v(r)}$ stands for a v(r)-dimensional vector with all elements equal to one. The definitions of the other elements in T^* are now obvious. Then T is equivalent with T^* with the k-th and last row deleted. Let T^* be partioned as $T^{*'} = [T_\eta, T'_\omega]$, with $T_\delta = [0, X_1, X_2, \dots, X_r, \dots, X_{K-1}, x_K]$. The subscript of T_η is motivated by the fact that this matrix is associated with the sufficient statistics for the η-parameters. The motivation for the subsript of T_ω is similar. Let $\{x \mid x_i = 1\}$ be the set of all possible response patterns with a response in category j of the i-th item and let the probabilities π_x in π have the same ordering as the response patterns in T_η. Then the ij-th element of the k-dimensional vector $T_\eta \pi$ is given by $\sum_{\{x \mid x_{ij}=1\}} \pi_x$. Using (14) it follows that

$$\sum_{\{x \mid x_{ij}=1\}} \pi_x = \sum_{r=0}^{K} v_r \varepsilon_{ij} \Gamma^{(i)}_{r-j} / \Gamma_r , \tag{17}$$

where $\Gamma^{(i)}_{r-j}$ is an elementary symmetric function of the order r-1 defined by

$$\varepsilon_{ij} \Gamma^{(i)}_{r-j} \triangleq \sum_{\{x \mid r(x)=r \text{ and } x_{ij}=1\}} \prod_{i=1}^{k} \varepsilon_{ij}^{x_{ij}} \tag{18}$$

and $\{x \mid r(x) = r \text{ and } x_{ij} = 1\}$ stands for the set of all possible response patterns with a response in category j of item I, resulting in sum score r.

A small example may clarify this assertion. Consider a test of three items with $m(i)=2$, and let $\boldsymbol{\pi}_r$ be the vector with elements $\pi_{\mathbf{x}}$, $\mathbf{x}\in\{\mathbf{x}\,|\,r(\mathbf{x})=r\}$. Then, for $r=2$, $X_2\boldsymbol{\pi}_2$ is given by

$$\begin{bmatrix} 1&1&0&0&0&0\\ 0&0&0&1&0&0\\ 1&0&1&0&0&0\\ 0&0&0&0&1&0\\ 0&1&1&0&0&0\\ 0&0&0&0&0&1 \end{bmatrix} \begin{bmatrix} \pi_{101000}\\ \pi_{100010}\\ \pi_{001010}\\ \pi_{010000}\\ \pi_{000100}\\ \pi_{000001} \end{bmatrix} = (\upsilon_2/\Gamma_2) \begin{bmatrix} \varepsilon_{11}\varepsilon_{21}+\varepsilon_{11}\varepsilon_{31}\\ \varepsilon_{11}\varepsilon_{21}+\varepsilon_{21}\varepsilon_{31}\\ \varepsilon_{11}\varepsilon_{31}+\varepsilon_{21}\varepsilon_{31}\\ \varepsilon_{12}\\ \varepsilon_{22}\\ \varepsilon_{32} \end{bmatrix} = \begin{bmatrix} \upsilon_2\varepsilon_{11}\,\Gamma_1^{(1)}/\Gamma_2\\ \upsilon_2\varepsilon_{21}\,\Gamma_1^{(2)}\,\Gamma_2\\ \upsilon_2\varepsilon_{31}\,\Gamma_1^{(3)}/\Gamma_2\\ \upsilon_2\varepsilon_{12}\,\Gamma_0^{(1)}/\Gamma_2\\ \upsilon_2\varepsilon_{22}\,\Gamma_0^{(2)}/\Gamma_2\\ \upsilon_2\varepsilon_{32}\,\Gamma_0^{(3)}/\Gamma_2 \end{bmatrix}$$

Summing these vectors over $r=0,\ldots,K$ results in a vector with expressions equivalent to (17). Returning to the derivation of the CML estimation equations, the equation $T_\eta \mathbf{n} = NT_\eta\boldsymbol{\pi}$ can be written in a more familiar form by introducing

$$s_{ij} \stackrel{d}{=} \sum_{\{\mathbf{x}\,|\,x_{ij}=1\}} n_{\mathbf{x}} \quad , \tag{19}$$

for $i=1,\ldots,k$ and $j=1,\ldots,m(i)$. Notice that s_{ij} is the number of persons responding in category j of item i. It can be verified that the K-dimensional vector $T_\eta\,\mathbf{n}$ has elements s_{ij}, by observing that the product of $\mathbf{n}$ with the ij-th row of T_η is equivalent to the right-hand sum of (19). In the same manner the elements of $T_\eta\boldsymbol{\pi}$ can be evaluated by applying (17). The vectors $T_\omega\mathbf{n}$ and $T_\omega\boldsymbol{\pi}$ can be evaluated analogously. Thus, the estimation equations $T\mathbf{n} = NT\boldsymbol{\pi}$ can also be written as

$$s_{ij} = N\sum_{r=0}^{K} \upsilon_r\varepsilon_{ij}\,\Gamma_{r-j}^{(i)}/\Gamma_r \quad , \tag{20}$$

for $i=1,\ldots,k$ and $j=1,\ldots,m(i)$, excluding $i=k$ and $j=m(k)$, and

$$n_r = N\upsilon_r \, , \tag{21}$$

for $r=0,\ldots,K-1$. Combining (20) and (21) results in the CML estimation equations

$$s_{ij} = \sum_{r=0}^{K} n_r\varepsilon_{ij}\,\Gamma_{r-j}^{(i)}/\Gamma_r \tag{22}$$

In the first section of this paper it could be seen that the rating scale model (Andrich, 1978a) and binomial trials model (Andrich, 1978b) could be derived from the partial credit model by imposing linear restrictions on the item parameters. In the sequel some other examples of the use of linear restrictions will be sketched. Generally, imposing linear restrictions is equivalent with introducing

$$\xi(\eta) = H\beta \tag{23}$$

with $\xi(\eta)' \stackrel{d}{=} (\eta_{11},\ldots,\eta_{km(k)-1})$, dimension($\beta$) $\leq$ dimension($\xi(\eta)$), and H of full column rank. The estimation equations for these models can easily be derived by observing that if

$$\Delta(\xi) \stackrel{d}{=} \partial \ln L(\xi|n)/\partial\xi \quad , \tag{24}$$

the CML estimation equations are given by

$$\partial \ln L(\xi|n)/\partial\beta = (\partial\xi/\partial\beta')\,\partial \ln L(\xi|n)/\partial\xi = H'\,\Delta(\xi) = 0 \quad . \tag{25}$$

This will be returned to in Section 5, where it will be shown that it is especially the combination of linear restrictions and incomplete designs that proves to be fruitful.

3. Testing the model.

The problem of evaluating model fit in IRT models is often solved within the well-established framework of the general multinomial model (see, for instance, Bock and Aitkin, 1981). This approach proceeds as follows. Let N_x be the number of persons with response pattern x. Further, let $\boldsymbol{N}$ be a vector of frequency counts with elements N_x for all $x \in \{x\}$, where $\{ x \}$ stands for the set of all possible response patterns. Then $\boldsymbol{N}$ has a multinomial distribution defined by N and $\boldsymbol{\pi}$, where N is the number of respondents and $\boldsymbol{\pi}$ a vector with as elements the probabilities π_x of the response patterns. Testing the model against a general multinomial alternative can by done by applying Pearson's X^2 test

$$X^2 = \sum_{\{x\}} \frac{(N_x - N\pi_x)^2}{N\pi_x} \tag{26}$$

or by using the asymptotically equivalent likelihood-ratio statistic

$$G^2 = 2 \sum_{\{x\}} N_x \ln \left(N_x / (N \pi_x) \right) . \tag{27}$$

If the probabilities in (26) and (27) are evaluated using BAN (best asymptotically normal) estimates, such as an ML estimate or a minimum χ^2 estimate, it can be shown (see, for instance, Bishop, Fienberg and Holland, 1975) that both statistics are asymptotically χ^2-distributed.

For reasonable long tests, this approach has two drawbacks, both related to the large number of possible response patterns. First, the expectation of $\boldsymbol{N}$ tends to have very small elements, and its realization, the vector of frequency counts $\mathbf{n}$, tends to have very small elements and elements equal to zero. In such cases it is often suggested to pool patterns to obtain expected frequencies which are sufficiently large. This pooling, however, is a function of the data itself, and the asymptotic distribution of a test statistic based on pooled data can hardly be derived. The second, and probably most serious drawback, is that interpreting the causes of a possible misfit is hampered by the aggregation level of the test: the influence of particular items on the outcome of the test as well as other possible causes of misfit cannot be identified.

Glas (1988) and Glas and Verhelst (1989) evade these problems by defining test statistics which are based on some linear function of N and have power against specific model viola-tions. Let $\mathbf{p}$ be defined by $\mathbf{p} \stackrel{d}{=} N/N$ and let $\mathbf{y}$ be defined by $\mathbf{y} \stackrel{d}{=} N^{1/2}(\mathbf{p}-\hat{\boldsymbol{\pi}})$, where $\hat{\boldsymbol{\pi}}$ stands for $\boldsymbol{\pi}$ evaluated using a BAN estimate of $\boldsymbol{\xi}$, the vector of all, say s, model parameters. It can easily be seen that (26) can be written as $\mathbf{y}'\hat{\mathbf{D}}_{\pi}^{-1}\mathbf{y}$, with $\mathbf{D}_{\pi}$ a $v \times v$ diagonal matrix of the elements $\pi_{\mathbf{x}}$, for all $\mathbf{x} \in \{\mathbf{x}\}$. Given certain regularity conditions, this statistic has an asymptotic χ^2-distribution with $v - s - 1$ degrees of freedom (see, for instance, Rao, 1973; Bishop, Fienberg and Holland, 1975). The aggregation level of this statistic is altered by defining the transformation $\mathbf{d} \stackrel{d}{=} \mathbf{X}\mathbf{y}$, where X is a $u \times v$ "matrix of contrasts" ($u < v$) of rank u and $\mathbf{d}$ a u-dimensional "vector of deviates". Generally, the class of statistics has the form

$$R = \mathbf{d}'\hat{\mathbf{W}}^{-1}\mathbf{d} \tag{28}$$

where $\hat{\mathbf{W}}$ is the so-called "matrix of weights", defined by $\mathbf{W} \stackrel{d}{=} \mathbf{X}\mathbf{D}_{\pi}\mathbf{X}'$, evaluated using a BAN estimate of $\boldsymbol{\xi}$. Let A be a $v \times s$ matrix defined by $\mathbf{A} \stackrel{d}{=} \mathbf{D}_{\pi}^{-1/2}(\partial\boldsymbol{\pi}/\partial\boldsymbol{\xi}')$, let $\mathbf{1}_v$ be a v-dimensional vector with all elements equal to one and let $\mathbf{c}$ be a u-dimensional vector of constants. Further, $M(\mathbf{D}_{\pi}^{1/2}\mathbf{X}')$ stands for the linear manifold spanned by the columns of $\mathbf{D}_{\pi}^{1/2}\mathbf{X}'$. Glas and Verhelst (1989) proved that R has an asymptotic χ^2-distribution with $u - s - 1$ degrees of freedom if

1. the columns of A belong to $M(\mathbf{D}_{\pi}^{1/2}\mathbf{X}')$, and
2. there exists a vector of constants $\mathbf{c}$ such that $\mathbf{X}'\mathbf{c} = \mathbf{1}_v$.

Notice that the number of degrees of freedom is equal to the number of deviates on which the test is based, minus the number of parameters to be estimated, minus one. Using these principles Glas and Verhelst (1989) introduced two tests for the partial credit model which can be used in a, so-called, marginal maximum likelihood framework. In the present paper a test for the partial credit model will be presented which applies in a CML framework. The test to be presented can be viewed as a generalization of, and an improvement upon, a test for the Rasch model for dichotomous items presented by Glas (1988). The nature of the improvement will be discussed later.

Let, for $r = 1,\ldots,K-1$, $i = 1,\ldots,k$ and $j = 1,\ldots,m(i)$, M_{rij} be the number of persons obtaining sum score r and responding in category j of item i. The counts M_{rij} associated with the scores $r = 0$ and $r = K$ will be considered later, because they are special in the

sense that there exists only one response pattern to obtain them and, as a consequence, for all i and j, $M_{r|j} = 0$ if $r = 0$ and $M_{rim(i)} = N_K$ if $r = K$. These are, however, not the only restrictions on the counts M_{rij}. If $r < j$, $M_{rij} = 0$, because it is not possible to respond in category j and obtain sum score $r < j$. In the same manner, it is not possible to respond in category j and obtain sum score $r > K - m(i) + j$. Therefore, these counts will be excluded, and only the counts M_{rij}, for $r = j,\ldots,K-m(i)+j$, will be considered.

For the construction of test statistics, the theoretical framework sketched above will be used. The starting point of the derivation is the multinomial model with probabilities π_x defined by (14). The expectation of M_{rij} can be derived by summing (14) over the set of all response patterns with $x_{ij} = 1$ resulting in sum score r, and multiplying by N. Using (17) it follows that

$$E(M_{rij}|\xi) = N \sum_{\{x \mid r(x)=r \text{ and } x_{ij}=1\}} \pi_x = N v_r \varepsilon_{ij} \Gamma^{(i)}_{r-j} / \Gamma_r \qquad (29)$$

and, hence,

$$E(M_{rij}|\hat{\xi}) = n_r \hat{\varepsilon}_{ij} \hat{\Gamma}^{(i)}_{r-j} / \hat{\Gamma}_r \quad . \qquad (30)$$

The fit of specific items to the model can be evaluated by, for $j = 1,\ldots,m(i)$ and $r = j,\ldots,K-m(i)+j$, inspecting the scaled deviates $d^{o}_{rij} \stackrel{d}{=} \left(M_{rij} - E(M_{rij}|\hat{\xi})\right) / \mathrm{var}(M_{rij}|\hat{\xi})^{1/2}$. The interpretation of the magnitude of d^{o}_{rij} may be helped by the fact that, if only one item, one category, and one score level are considered, and no parameters have to be estimated, d^{o}_{rij} is a standardized biniomial variable. Squaring and summing d^{o}_{rij} over the appropriate range of sum scores yields an index of item fit that would be approximately χ^2-distributed, if the assumptions given should hold. They do, of course, not hold, but, in conjunction with a formal test of model fit based on the difference between the observed and expected values of the counts M_{rij}, the indices d^{o}_{rij} can be used for identifying items that have contributed most to a possible lack of model fit.
A formal model test based on these deviates can be constructed as follows. First an elementary function $\Gamma^{(i,i')}_{r-j-j'}$ must be defined. Let $\{x \mid r(x)=r \text{ and } x_{ij}=1 \text{ and } x_{i'j'}=1\}$ stand for the set of all possible response patterns leading to sum score r ($r \geq j+j'$) with $x_{ij}=1$ and $x_{i'j'}=1$. Then

$$\varepsilon_{ij}\varepsilon_{i'j'}\Gamma^{(i,i')}_{r-j-j'} \overset{d}{=} \sum_{\{x \mid r(x)=r \text{ and } x_{ij}=1 \text{ and } x_{i'j'}=1\}} \prod_{i,j} \varepsilon_{ij}^{x_{ij}} \quad . \tag{31}$$

As already mentioned elementary functions of order less than zero are supposed to be zero. This definition will be used in the following theorem.

Theorem 1. For $r = 1,\ldots,K-1$, let $\mathbf{d}_r$ be a vector with elements defined by

$N^{1/2} d_{rij} \overset{d}{=} M_{rij} - E(M_{rij} \mid \hat{\boldsymbol{\xi}})$, for $i = 1,\ldots,k$ and $j = \max(1, r+m(i)-K),\ldots,\min(m(i), r)$.

The dimension of $\mathbf{d}_r$ is $e_r \overset{d}{=} \sum_i \left[\min(m(i), r) - \max(1, r+m(i)-K)+1\right]$.

Let W_r be an $e_r \times e_r$ matrix. If d_{rij} is the p-th element of $\mathbf{d}_r$, $d_{rij'}$ is the p'-th element of $\mathbf{d}_r$ $(j \neq j')$ and $d_{ri'j''}$ is the p''-th element of $\mathbf{d}_r$ $(i \neq i')$, the elements $W_r(p,p)$, $W_r(p,p')$ and $W_r(p,p'')$ of W_r are defined by

$$W_r(p,p) \overset{d}{=} \upsilon_r \, \varepsilon_{ij} \Gamma^{(i)}_{r-j} / \Gamma_r$$

$$W_r(p,p') \overset{d}{=} 0 \quad \text{and}$$

$$W_r(p,p'') \overset{d}{=} \upsilon_r \, \varepsilon_{ij}\varepsilon_{i'j''} \Gamma^{(i,i')}_{r-j-j''} / \Gamma_r \quad .$$

Then $$R_{1c} \overset{d}{=} \sum_{r=1}^{K-1} \mathbf{d}'_r \hat{W}_r^{-1} \mathbf{d}_r \tag{32}$$

has an asymptotic χ^2-distribution with $\sum_{r=1}^{K-1} e_r - 2K + 2$ degrees of freedom.

The computation of the number of degrees of freedom will be discussed in the next section, where the proof of the Theorem is given. The following corollary can be used for models that can be derived from the unidimensional Rasch model for polytomous items by imposing linear restrictions on the item parameters.

Let $\boldsymbol{\xi}(\boldsymbol{\eta}) \overset{d}{=} (-\eta_{11},\ldots,-\eta_{ij},\ldots,-\eta_{k(m(k)-1)})$.

Corollary 1. If the item parameters are subject to linear restrictions of the form $\boldsymbol{\xi}(\boldsymbol{\eta}) = H\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ an s'-dimensional vector of parameters and H of rank s', the test statistic R_{1c} defined in Theorem 1 has an asymptotic χ^2-distribution with $\sum_{r=1}^{K-1} e_r - s' - K + 1$ degrees of freedom.

The test statistic defined in Theorem 1 is based on partitioning respondents in score groups. If the number of possible scores is large, it is often practical to create a partitioning with less classes. This can be done by combining score groups into a new subgroup. One reason for doing this is the fact that if the number of possible scores is large, identification of the causes of misfit is hampered by the large number of deviates to be inspected. Another reason may be that for certain score levels, the expectation of the counts may be very low, which may invalidate the asymptotic results on which the derivation of the statistic is based.

Suppose that the total sample of respondents is partitioned into G ($G > 1$) subgroups and that, for $g = 1,\ldots,G$, $l(g)$ is the smallest and $u(g)$ is the largest score of the respondents included in subgroup g. Of course, $l(1) = 1$ and $u(G) = K-1$. Further, it will prove convenient that the largest score included in the first subgroup is greater than, or equal to, $\max(m(i))$, hat is $u(1) \geq \max(m(i))$, and the smallest score included in the last subgroup is smaller than or equal to $\min(K-m(i)+1)$, that is $l(G) \leq \min(K-m(i)+1)$. The reason for introducing the last two assumptions will be commented upon after the following theorem.

Theorem 2. Let $\mathbf{d}_{(g)}$, $g = 1,\ldots,G$, be K-dimensional vectors of deviates with elements defined by $N^{1/2} d_{(g)ij} \stackrel{d}{=} \sum_{r=l(g)}^{u(g)} M_{rij} - E(M_{rij} | \hat{\xi})$, for $i = 1,\ldots,k$.

Let the elements of G matrices of weights $W_{(g)}$ be defined by

$$W_{(g)}(ij,ij) \stackrel{d}{=} \sum_{r=l(g)}^{u(g)} v_r \varepsilon_{ij} \Gamma_{r-j}^{(i)} / \Gamma_r - \sum_{r=l(g)+1}^{u(g)} v_r \left[\varepsilon_{ij} \Gamma_{r-j}^{(i)} / \Gamma_r \right]^2 ,$$

$$W_{(g)}(ij,ij') \stackrel{d}{=} - \sum_{r=l(g)+1}^{u(g)} v_r \left[\varepsilon_{ij} \Gamma_{r-j}^{(i)} / \Gamma_r \right] \left[\varepsilon_{ij'} \Gamma_{r-j'}^{(i)} / \Gamma_r \right] , \text{ with } j \neq j',$$

and, if $i \neq i'$,

$$W_{(g)}(ij,i'j') \stackrel{d}{=} \sum_{r=l(g)}^{u(g)} v_r \varepsilon_{ij} \varepsilon_{i'j'} \Gamma_{r-j-j'}^{(i,i')} / \Gamma_r - \sum_{r=l(g)+1}^{u(g)} v_r \left[\varepsilon_{ij} \Gamma_{r-j}^{(i)} / \Gamma_r \right] \left[\varepsilon_{i'j'} \Gamma_{r-j'}^{(i')} / \Gamma_r \right] .$$

Then
$$R_{1c} \stackrel{d}{=} \sum_{g=1}^{G} \mathbf{d}'_{(g)} \hat{W}_{(g)}^{-1} \mathbf{d}_{(g)} \qquad (33)$$

has an asymptotic χ^2-distribution with $(G-1)(K-1)$ degrees of freedom.

The degrees of freedom will be explained in the next section, where the proof of the theorem is given.

The reason for the restriction $u(1) \geq \max(m(i))$ is given by the observation that if $u(1)$ were chosen such that some item i had a category $j > u(1)$, the diagonal element $W_{(g)}(ij,ij)$ would be equal to zero, because $\Gamma^{(i)}_{r-j} = 0$ if $j > r$. As a result $\hat{W}_{(g)}$ would not be invertible. Notice, by the way, that in this case $d_{(g)ij}$ is also equal to zero.
In the same manner, it can also be verified that violating the restriction $I(G) \leq \min(K-m(i)+1)$ results in elements $d_{(g)ij}$ and $W_{(g)}(ij,ij)$ equal to zero.
Again, the version of the model test defined by (33) can be applied to the case of linear restrictions on the item parameters.

Corollary 2. If the item parameters are subject to linear restrictions of the form $\xi(\eta) = H\beta$, with β an s'-dimensional vector of parameters and H has rank s', the test statistic R_{1C} defined in Theorem 2 has an asymptotic χ^2-distribution with $G(K-1) - s'$ degrees of freedom.

In the next section the proof of Theorems 1 and 2 and their corollaries will be given.

4. The derivation of the asymptotic distribution of R_{1c}.

In the present section, Theorem 2 will be proved first. Then it will be sketched what alterations have to be made to prove Theorem 1. In either proof, the general framework developed in Glas and Verhelst (1989) will be used. The general framework relates to a multinomial model with response patterns as categories, so let $\mathbf{p}$ be a v-dimensional vector with elements $p_{\mathbf{x}}$, for all $\mathbf{x} \in \{\mathbf{x}\}$, where $p_{\mathbf{x}}$ is the observed proportion of persons with response pattern $\mathbf{x}$. Further, let $\boldsymbol{\pi}$ be a v-dimensional vector with elements $\pi_{\mathbf{x}}$, for all $\mathbf{x} \in \{\mathbf{x}\}$, where $\pi_{\mathbf{x}}$ is the probability of observing response pattern $\mathbf{x}$, defined by (15). Consider a test statistic defined by $R_{1c} \stackrel{d}{=} \mathbf{d}'\hat{\mathbf{W}}^{-1}\mathbf{d}$, where $\mathbf{d} \stackrel{d}{=} N^{1/2}\mathbf{X}(\mathbf{p} - \hat{\boldsymbol{\pi}})$ and $\mathbf{W} \stackrel{d}{=} \mathbf{X}\mathbf{D}_{\boldsymbol{\pi}}\mathbf{X}'$. The matrix of contrasts X is defined by

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}' & & & \\ & \mathbf{X}_{(1)} & & \mathbf{0} \\ & & \mathbf{X}_{(g)} & \\ & \mathbf{0} & & \mathbf{X}_{(G)} \\ & & & & \mathbf{1}' \end{bmatrix}, \tag{34}$$

where $\mathbf{X}_{(g)}$ is defined by

$$\mathbf{X}_{(g)} \stackrel{d}{=} \begin{bmatrix} \mathbf{X}_{l(g)} & \mathbf{X}_{l(g)+1} & \cdots & \mathbf{X}_r & \cdots & \mathbf{X}_{u(g)} \\ \mathbf{0}' & \mathbf{1}' & \cdots & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{1}' & \cdots & \mathbf{0} \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{0}' & \cdots & \mathbf{1}' \end{bmatrix}, \tag{35}$$

with $\mathbf{X}_r$ a $K \times v_r$ matrix with as columns all $\mathbf{x} \in \{\mathbf{x} \mid r(\mathbf{x}) = r\}$. Further, $\mathbf{1}$ is a vector with all elements equal to one, and $\mathbf{0}$ is a vector with all elements equal to zero. The dimensions of these vectors are equal to the number of columns in the matrices $\mathbf{X}_r$, $r = l(g),\ldots,u(g)$, above them. In the sequel it will prove convenient to partition $\mathbf{X}_{(g)}$ as

$$\mathbf{X}_{(g)} = \begin{bmatrix} \mathbf{Z}_g \\ \mathbf{Y}_g \end{bmatrix}, \tag{36}$$

with $\mathbf{Z}_g \stackrel{d}{=} \begin{bmatrix} \mathbf{X}_{l(g)} & \mathbf{X}_{l(g)+1} & \cdots & \mathbf{X}_r & \cdots & \mathbf{X}_{u(g)} \end{bmatrix}$.

First $\mathbf{A} \stackrel{d}{=} \mathbf{D}_{\boldsymbol{\pi}}^{-1/2}\partial\boldsymbol{\pi}/\partial\boldsymbol{\xi}'$ must be derived. Let the matrix $\mathbf{T}^*$ be defined as

$$T^{*} = \begin{bmatrix} 0 & X_1 & X_2 & \dots & X_r & \dots & X_{K-1} & x_K \\ 1 & 0' & 0' & \dots & 0' & \dots & 0' & 0 \\ 0 & 1' & 0' & \dots & 0' & \dots & 0' & 0 \\ 0 & 0' & 1' & \dots & 0' & \dots & 0' & 0 \\ & & & & & & & \\ 0 & 0' & 0' & \dots & 1' & \dots & 0' & 0 \\ & & & & & & & \\ 0 & 0' & 0' & \dots & 0' & \dots & 1' & 0 \\ 0 & 0' & 0' & \dots & 0' & \dots & 0' & 1 \end{bmatrix}, \tag{37}$$

where X_r is a $K \times v_r$ matrix with as columns all $x \in \{x \mid r(x) = r\}$, and x_K is the pattern with all responses in the highest possible category.

Since the model belongs to an exponential familly, it follows from Lemma 1, that $\partial\pi/\partial\xi' = D_\pi T' - \pi\pi' T'$, where T is equal to T^{*}, with the K-th and last row deleted. The asymptotic distribution of the test can now be derived by checking conditions 1 and 2. In the following lemma, T^{*} will be partitioned

$$T^{*} = \begin{bmatrix} T_\eta \\ T_\omega \end{bmatrix}, \tag{38}$$

such that $T_\eta \overset{d}{=} [0, X_1, X_2, \dots, X_r, \dots, X_{K-1} x_K] = [0, Z_1, \dots, Z_g, \dots, Z_G, x_K]$.

Lemma 2. The columns of A belong to $M(D_\pi^{1/2}X')$ and there exists a vector $\mathbf{c}$ such that $X'\mathbf{c} = 1_v$.

Proof. It will first be proved that the columns of $D_\pi^{1/2}T'$ belong to $M(D_\pi^{1/2}X')$ by showing that the columns of $D_\pi^{1/2}T^{*\prime}$ belong to this manifold. From

$$D_\pi^{1/2}T'_\eta = D_\pi^{1/2}\begin{bmatrix} \mathbf{0}' \\ Z'_1 \\ Z'_g \\ Z_G \\ \mathbf{x}'_K \end{bmatrix} \quad \text{and} \quad M\left(D_\pi^{1/2}\begin{bmatrix} 1 & & & \\ & Z'_1 & & 0 \\ & & Z'_g & \\ & 0 & & Z'_G \\ & & & & 1 \end{bmatrix}\right) \subset M(D_\pi^{1/2}X') .$$

it follows that the columns of $D_\pi^{1/2}T'_\eta$ belong to $M(D_\pi^{1/2}X')$. That the columns of $D_\pi^{1/2}T'_\omega$ also belong to this manifold can be verified in the following manner. Consider (35). A vector with all elements equal to one can be constructed from $X'_{(g)}$ as $X'_{(g)}\mathbf{c}_g$, with the first K elements of $\mathbf{c}_g$ equal to the elements of the K-dimensional vector $(1/l(g))\,(1,2,\ldots,m_1,\ldots,1,2,\ldots,m_i,\ldots,1,2,\ldots,m_k)$ and the remaining elements equal to $(l(g)-r)/l(g)$, for $r = l(g)+1,\ldots,u(g)$. Applying this procedure to all matrices in X, it can be shown that if $\mathbf{c}' = (1,\mathbf{c}'_1,\ldots,\mathbf{c}'_g,\ldots,\mathbf{c}_G,1)$, $X'\mathbf{c} = 1_v$.
From the last observation it also follows that the columns of $D_\pi^{1/2}T'_\omega$ belong to $M(D_\pi^{1/2}X')$. Finally, $D_\pi^{1/2}X'\mathbf{c} = D_\pi^{1/2}1_v = \pi^{1/2}$, and so $\pi^{1/2}\pi'T'$ belongs to $M(D_\pi X')$. □

Next, it will be shown that the test statistic defined by $R_{1c} = \mathbf{d}'\hat{W}^{-1}\mathbf{d}$ is equivalent with the statistic defined in Theorem 2.

Let $\mathbf{p}'_{(g)} \overset{d}{=} (\mathbf{p}'_{l(g)},\ldots,\mathbf{p}'_r,\ldots,\mathbf{p}'_{u(g)})$, where $\mathbf{p}_r$ has elements $p_x = N_x/N$, for all $x \in \{x \mid r(x)=r\}$.
In the same manner, $\boldsymbol{\pi}'_{(g)} \overset{d}{=} (\boldsymbol{\pi}'_{l(g)},\ldots,\boldsymbol{\pi}'_r,\ldots,\boldsymbol{\pi}'_{u(g)})$, where $\boldsymbol{\pi}_r$ has elements π_x, for all $x \in \{x \mid r(x)=r\}$. Further, $D_{\pi(g)}$ is the diagonal matrix of the elements of $\boldsymbol{\pi}_{(g)}$. From the structure of X, it follows that $\mathbf{d}'\hat{W}^{-1}\mathbf{d}$ can also be written as $\sum_{g=1}^{G} \mathbf{d}'_{*g}\hat{W}_{*g}^{-1}\mathbf{d}_{*g}$, with $W_{*g} = X_{(g)}D_{\pi(g)}X'_{(g)}$ and $\mathbf{d}_{*g} = N^{1/2}X_{(g)}(\mathbf{p}_{(g)} - \hat{\boldsymbol{\pi}}_{(g)})$.
The elements of $NY_{(g)}(\mathbf{p}_{(g)} - \hat{\boldsymbol{\pi}}_{(g)})$ are given by $\left[N_r - E(N_r \mid \hat{\boldsymbol{\xi}})\right]$, for $(l(g)+1) \le r \le u(g)$, and from $N_r = n_r$ and $\hat{\upsilon}_r = n_r/N$ it follows that these elements are all equal to zero. Therefore, $\mathbf{d}_{*g}$ can be given by $\mathbf{d}'_{*g} = (\mathbf{d}'_{(g)},\mathbf{0})$, with $\mathbf{d}_{(g)} = N^{1/2}Z_{(g)}(\mathbf{p}_{(g)} - \hat{\boldsymbol{\pi}}_{(g)})$. Let the matrix W_{*g} be partitioned

$$W_{*g} = \begin{bmatrix} W_{*g11} & W_{*g12} \\ W_{*g21} & W_{*g22} \end{bmatrix} = \begin{bmatrix} Z_g D_{\pi(g)} Z'_g & Z_g D_{\pi(g)} Y'_g \\ Y_g D_{\pi(g)} Z'_g & Y_g D_{\pi(g)} Y'_g \end{bmatrix} \tag{39}$$

As a concequence of the fact that $\mathbf{d}_g$ can be given by $\mathbf{d}'_g = (\mathbf{d}'_{(g)}, \mathbf{0})$,

$$R_{1c} = \sum_{g=1}^{G} \mathbf{d}'_{(g)} \hat{W}^{-1}_{(g)} \mathbf{d}_{(g)} = \sum_{g=1}^{G} \mathbf{d}'_{(g)} \left[\hat{W}_{*g11} - \hat{W}_{*g12} \hat{W}^{-1}_{*g22} \hat{W}_{*g21} \right]^{-1} \mathbf{d}_{(g)} \tag{40}$$

and since W_{*g22} is a diagonal matrix of the elements v_r, for $r = l(g)+1,\ldots,u(g)$, it can be verified that the test statistic defined in the present section is equivalent with the statistic defined in Theorem 2. The number of degrees of freedom of the test can be computed by counting the number of deviates on which the test is based, which is equal to the dimension of $(1, \mathbf{d}'_{*1}, \ldots, \mathbf{d}'_{*g}, \ldots, \mathbf{d}'_{*G}, 1)$, that is $GK + 2 + (K - 1 - G)$, minus the dimension of $\boldsymbol{\xi}$, which is $2K - 1$, minus one. Thus, the number of degrees of freedom is equal to $(G-1)(K-1)$. If the item parameters are a linear function of an s'-dimensional parameter vector $\boldsymbol{\beta}$, the number of degrees of freedom is equal to $GK + 2 + (K-1-G)$ minus s', K, and one, so in this case, the resulting number of degrees of freedom is equal to $G(K-1) - s'$.

Next, the proof of Theorem 1 will be sketched. In this case, $K-1$ subgroups are made and, for every subgroup g, $g = 1,\ldots,G$, $l(g) = u(g)$. As a result, the matrix of contrasts has the form

$$X = \begin{bmatrix} X_1 & & & 0 \\ & X_2 & & \\ & & X_r & \\ 0 & & & X_{K-1} \\ & & & 1 \end{bmatrix}. \tag{41}$$

If N_r has elements N_x, $x \in \{x \mid r(x) = r\}$, $X_r N_r$ will have elements M_{rij}, for $i = 1,\ldots,k$ and $j = 1,\ldots,m(i)$. In the previous section, however, it was argued that for certain combinations of i,j and r, M_{rij} is systematically equal to zero. The rows of X_r associated with these elements are equal to the zero-vector, and, as a result, X will be of incomplete rank, which obstructs using the theory sketched in section 3. Therefore, a matrix X^*_r will be introduced, which can be derived from X_r by removing all rows which

are associated with a combination of the indices i and j for which $r < j$ or $r > K-m(i)+j$. If X_r has $\sum_i m(i)$ rows, X_r^* will have e_r rows. Let the matrix of contrasts X^* be defined by

$$X^* \triangleq \begin{bmatrix} 1 & & & & \\ & X_1^* & & & \\ & & X_2^* & & 0 \\ & & & \ddots & \\ & 0 & & X_r^* & \\ & & & & \ddots \\ & & & & X_{K-1}^* \\ & & & & & 1 \end{bmatrix}. \qquad (42)$$

Notice that X' and $X^{*'}$ have the same column space, since the latter is derived from the former by removing zero-vectors. Therefore, proving that the columns of A belong to $M(D^{1/2}X_\pi^{*'})$ is equivalent with proving that they belong to $M(D_\pi^{1/2}X')$, and from the existence of $\mathbf{c}$ such that $X^{*'}\mathbf{c} = 1_v$ it follows that $X'\mathbf{c} = 1_v$. However, the last two proofs are given in Lemma 2, if $G = K-1$ is chosen.

The number of degrees of freedom is computed as follows. Notice that X^* has $2 + \sum_{r=1}^{K-1} e_r$ rows, so the number of deviates on which the testing procedure is based is equal to $2 + \sum_{r=1}^{K-1} e_r$. The number of parameters to be estimated is $2K-1$. As a result, the number of degrees of freedom of this version of the R_{1c} test is $\sum_{r=1}^{K-1} e_r - 2K + 2$.

In Glas (1989) it was shown that if a statistic of the form defined by (28) has an asymptotic χ^2-distribution for some parameterized multinomial model, the statistic will also be χ^2-distributed for models that can be derived by imposing linear restrictions on the parameters of the more general model. This proves corollaries 1 and 2.

5. *Some generalizations of the estimation procedures.*

The present section will be mainly devoted to generalizing the estimation procedures for the unidimensional Rasch model for polytomous data, or partial credit model, to incomplete designs and linear restrictions on the parameters. Let the items be indexed $i = 1,...,k$ and the respondents by $n = 1,...,N$. A test administration design consists of T tests, indexed $t = 1,...,T$, and every test is identified by the pair $\left[\{n\}_t, \{i\}_t\right]$, where $\{n\}_t$ is the set of the indices of the persons taking test t, and $\{i\}_t$ is the set of the indices of the items in test t. It is assumed that for every two tests t and t', $\{n\}_t \cap \{n\}_{t'} = \emptyset$. Further, the design vector $\mathbf{a}_t'$ has elements a_{ti} for $i = 1,...,k$, where $a_{ti} = 1$ if item $i \in \{i\}_t$ and $a_{ti} = 0$ if this is not the case. First it will be assumed that the tests are linked via common items, which means that, for every two tests indexed t and t', there exists a sequence of indices $z_1, z_2, ..., z_h$, such that $\mathbf{a}_t'\mathbf{a}_{z_1} > 0$, $\mathbf{a}_{z_1}'\mathbf{a}_{z_2} > 0, ..., \mathbf{a}_{z_h}'\mathbf{a}_{t'} > 0$. In the sequel it will be shown that in some instances this assumption may be dropped.

Derivation of estimation equations can be done using the fact that, due to experimental independence, the complete log-likelihood function can be written as the sum of the log-likelihood functions of the tests. Applying this principle to the CML estimation equations (22), the CML estimation equations in an incomplete design are given by

$$s_{ij} = \sum_{t=1}^{T} a_{ti} \sum_{r=0}^{K_t} n_{tr} \frac{\varepsilon_{ij}\Gamma^{(i)}_{t(r-j)}}{\Gamma_{tr}}, \tag{43}$$

for $i = 1,...,k$ and $j = 1,...m(i)$, with s_{ij} the observed number of responses in category j of item i, K_t the maximum score that can be obtained on test t, and n_{tr} the number of persons obtaining score r on test t. Further, Γ_{tr} is an elementary function of order r, of the parameters ε_{ij}, $i \in \{i\}_t$ and $j = 1,...,m(i)$, and $\Gamma^{(i)}_{t(r-j)}$ is an elementary function of order $r - j$ of the parameters ε_{ij}, $i \in \{i\}_t$ and $j = 1,...,m(i)$, where all parameters associated with item i have been set equal to zero.

As in the case of a complete design, the item parameters may be subject to linear restrictions of the form $\xi(\eta) = H\beta$, with $\xi(\eta)' = (-\eta_{11},\ldots,-\eta_{ij},\ldots,-\eta_{k(m(k)-1)})$, β an s-dimensional vector of parameters and H a matrix of full column rank. Again, the CML estimation equations for β are given by $H'\Delta(\xi(\eta)) = 0$, where $\Delta(\xi(\eta))$ is the vector of first-order derivatives of the complete log-likelihood function with respect to $\xi(\eta)$.

It may be interesting to notice that if linear restrictions are imposed, the design needs no longer be necessarily linked. Consider an example of a design with two tests, where, for $t = 1$ and $t = 2$, $\xi(\eta)^{(t)}$ are item parameter vectors which have no elements in common, i.e., $\xi(\eta)$ can be written as $\xi(\eta)' = (\xi(\eta)^{(1)}, \xi(\eta)^{(2)})'$. So it is assumed that the tests are not linked. Let

$$\begin{bmatrix} \xi(\eta)^{(1)} \\ \xi(\eta)^{(2)} \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \end{bmatrix} . \tag{44}$$

Then the estimation equations are given by

$$\begin{bmatrix} H_{11}' & H_{21}' \\ H_{12}' & H_{22}' \end{bmatrix} \begin{bmatrix} \Delta(\xi(\eta))^{(1)} \\ \Delta(\xi(\eta))^{(2)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} , \tag{45}$$

with $\Delta(\xi(\eta)^{(t)})$ the vector of first order derivatives of the log-likelihood function with respect to $\xi(\eta)^{(t)}$. Since the design is not linked by common items, the two systems of linear equations, $\Delta(\xi(\eta)^{(1)}) = 0$ and $\Delta(\xi(\eta)^{(2)}) = 0$, are independent. If $H_{12} = 0$ and $H_{21} = 0$, the system (45) is equivalent with the two independent systems $H_{11}'\Delta(\xi(\eta)^{(1)}) = 0$ and $H_{22}'\Delta(\xi(\eta)^{(2)}) = 0$. So, generally speaking, it is a necessary condition for the existence of a unique solution of (45), that there does not exist a permutation of the elements of β and an associated permutation of the rows of H, such that matrices H_{12} and H_{21} are obtained that are both equal to zero. Derivation of sufficient conditions for the existence of a solution of the estimation equations for β, however, is beyond the scope of the present paper, for this topic one is referred to Fischer(1981,1983).

The present section will be concluded with an example of the possibilities opened up by combination of an incomplete design and linear restrictions on the parameters. Suppose N persons are confronted with k' open-ended questions. The responses are judged by two referees. For $i=1,\ldots,k'$, let X_{ni} be the judgement of the first, and $X_{n(k'+i)}$ the judgement of the second referee with respect to the response of person n on item i. Clearly, X_{ni} and $X_{n(k'+i)}$ are not independent, for they both depend on the same response and the derivation of a likelihood-function is hampered by the fact that the product-rule cannot be used. To circumvent this problem, the respondents are divided into two groups. For the first group, only the judgements of the first referee are included in the analysis and for the second group, only the judgements of the second referee are considered. This results in the design displayed in Figure 1. The data which are not included in the study can be used for a cross-validation.
The differences in the judgements of the two referees can now be modeled by imposing restrictions on the item parameters such as

(1) $\eta_{(k'+i)j} = \eta_{ij} + \tau$, which is a model without interaction between judges and items
(2) $\eta_{(k'+i)j} = \eta_{ij} + \tau_i$, which is a model with interaction between judges and items
(3) $\eta_{(k'+i)j} = \eta_{ij} + \tau_j$, which is a model with interaction between judges and categories
(4) $\eta_{(k'+i)j} = \eta_{ij} + \tau_{ij}$, which is a model with interaction between judges and combinations of items and categories.

The models (1), (2) and (3) can be estimated both by CML and MML estimation procedures. For the MML estimation and testing procedure one is referred to Glas and Verhelst (1989). Using Fischer (1983, Theorem 3, p.16), it can be easily verified that model (4) is not identified in a CML framework. In an MML framework, model (4) is a reparametrization of the Rasch

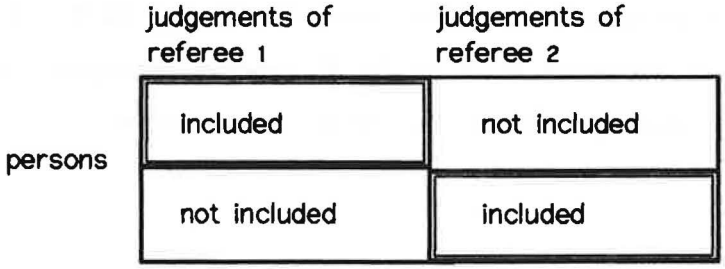

Figure 1. Data collection design for studying interjudge reliability.

model rather than a restricted version. Further, in an MML framework, the referee effects in case (1) could also be modeled by the population parameters, introducing $\mu_1 = \mu$ as the mean of the ability distribution of the first group and $\mu_2 = \mu - \tau$ as the mean of the ability distribution of the second group.
This strategy for analyzing referee effects can, of course, be generalized in various ways. One may, for instance, think of a design with more than two judges, or of interaction effects between judges and subgroups of respondents. With respect to this last example, one may, for instance, think of interaction effects between sex or race of the referee and the respondent. These generalizations will, however, not be worked out in detail in the present paper.

6. Some generalizations of the testing procedure.

To be able to use the theory for the construction of test statistics sketched in Section 3, the model for incomplete designs have to be brought within the framework of the multinomial model. For the conditional model defined by (14), this is accomplished by introducing the assumption that, if N_{tr} is the number of persons obtaining a sum score r, N_{tr} for $t = 1,\ldots,T$ and $r = 0,\ldots,K_t$ has a multinomial distribution characterized by N and the probabilities υ_{tr}, $t = 1,\ldots,T$ and $r = 0,\ldots,K_t$. The probability of a response pattern on test t can now be given by

$$\pi_{tx} \stackrel{d}{=} \upsilon_{tr} \prod_{i \epsilon \{x\}_t} \prod_{j} \varepsilon_{ij}^{x_{ij}} / \Gamma_{tr} \quad . \tag{46}$$

Let K be defined $K \stackrel{d}{=} (\sum_{i=1}^{k} m_i)$. The number of parameters to be estimated is the sum of K-1 item parameters and $(\sum_t K_t) + T - 1$ parameters associated with the distribution of N_{tr}, that is, $(\sum_t K_t) + T + K - 2$ parameters have to be estimated. This will be used for computing the degrees of freedom of the test statistics.

With the model brought within the framework of parameterized multinomial models, the method for constructing model tests sketched in section 3 can be applied. The generalization of the R_{1c} test to incomplete designs is accomplished by constructing a matrix of contrasts X, which has the form

$$X = \begin{bmatrix} X_1 & & 0 \\ & X_t & \\ 0 & & X_T \end{bmatrix} , \qquad (47)$$

where X_t is the usual matrix for testing a certain contrast in the case of one test. As a consequece of the structure of this super-matrix, the complete model test will be a sum of the statistics of the separate subgroups in the design. Thus, the R_{1c} statistic is generalized

$$R_{1c} \overset{d}{\cong} \sum_{t=1}^{T} R_{1c}^{(t)} , \qquad (48)$$

with $R_{1c}^{(t)}$ the statistic for the separate groups. Given the structure of the matrix of contrasts, the proof that the statistic defined above has an asymptotic χ^2-distribution is easily derived from the analogous proof for a complete design.

The number of degrees of freedom is computed as follows. First assume that no linear restrictions are imposed. Let the statistic be computed using G_t subgroups in test t. For every test t, the number of deviates is equal to $2 + G_t K_t + (K_t - 1) - G_t$, so the total number of deviates is $T + \sum_t G_t(K_t - 1) + \sum_t K_t$. Subtracting one plus the number of parameters to be estimated, which is $(\sum_t K_t) + T + K - 1$, results in $\sum_t G_t(K_t - 1) - (K - 1)$ degrees of freedom. In the same manner it can also be derived that the number of degrees of freedom is $\sum_t G_t(K_t - 1) - s'$, if s' linear functions of the item parameters are estimated.

7. Some examples.

A dichotomous item is defined to be biased if, for a given level of ability, the probability of a correct response differs over groups (Mellenbergh, 1982,1983). So though items may differ in difficulty and groups may differ in their ability to solve the item, but that does not define item bias. An item is only considered biased when it differs in difficulty between subjects of identical ability. The generalization to polytomous items is straightforward. A polytomous item can be considered biased if the set of probabilities of scoring in the various categories of an item, for a given ability level, differs between groups. Several techniques for detecting item bias have been proposed, all based on evaluating the differences response probabilities between groups conditional on some measure of ability. The most generally used technique is based on the Mantel-Haenszel statistic (Holland and Thayer, 1988), others are based on loglinear models (Kok, Mellenbergh and van der Flier, 1985) or on IRT models (Hambleton and Rogers, 1989). The advocates of the Mantel-Haenszel and loglinear approach evaluate item difficulty conditional on sum scores. It is well-known that adopting the sum score as a sufficient statistic for ability is equivalent with adopting the Rasch model (Fischer, 1984). The adherents of these approaches do not subscribe to this implication and do not actually use the Rasch model. Application of IRT to the problem of detecting item bias, on the other hand, suffers from the poor mathematical foundation of test statistics. An exception is the technique proposed by Kelderman (1989), which is based on a so-called loglinear IRT model. One of the drawbacks of this approach is that it is difficult to use on larger tests. Therefore the present author will suggest an alternative approach based on the theory presented above. Combination of the method presented here and Kelderman's techniques remains a topic of further study. Two examples will be given, the first pertains to dichotomous items, the second to polytomous items.

The first example is the 1990 examination in reading comprehension in English for the Dutch lower general secundary education, the MAVO-D level. The technique has also been applied to similar examinations of French and German, and to other educational levels (MAVO-C, HAVO and VWO), all with results analogous to the results presented. The analyses were carried out using a sample of 1000 boys and 1000 girls from the complete examination population. The examination was made up of 50 dichotomously scored multiple choice items. The objective of the procedure is to detect items for which

the probability of eliciting a correct response is higher or lower than predicted from the estimated ability level, that is the score level, of either the boys or the girls. Essentially the procedure consists of two stages: (1) identifying Rasch-homogeneous subscales, and (2) evaluating the differences in response probabilities between boys and girls in homogeneous ability groups.

In order not to let item bias interfere with searching for Rasch-homogeneous subscales, only the data of one of the two groups is used at first. In the present case the choice of a group is completely arbitrary, and the girls were chosen. In other instances one may start with a so-called reference group, and examine the responses of the so-called focal group in the second stage of the procedure.

Starting with the complete examination, item parameters were estimated and the R_1 statistic was computed. Next, items with the largest fit indices were removed and the process was repeated until the R_1 statistic was no longer significant. In the same manner the next Rasch scale was identified by repeating the process using the removed items. Three subscales of 21, 14 and 12 items, respectively, were identified, three items could not be categorized. In Table 1 the evaluation of model fit for the largest subscale is summarized.

Insert Table 1 about here

The total sample of female examinees was divided into six score groups of approximately the same number of respondents. In the rows of Table 1 the scaled deviates of the items are given. If no parameters had to be estimated these would be standard normal deviates. Squaring and summing scaled deviates over subgroups results in an index of item fit. Again if no parameters had to be estimated, the index would be chi-square distributed with six degrees of freedom. Of course, the assumption does not hold, but the fit indices serve their purpose in identifying the relative contribution of items to the R_1 statistic. The value of the R_1 statistic is given at the bottom of the table, together with its degrees of freedom and its probability.

Insert Table 2 about here

Next, the fit of the subscale was evaluated using the boys' data. The results are given in Table 2. Inspecting the value of the R_1 statistic at the bottom of the table reveals that the items fit a Rasch-scale, yet the fit is less perfect than with the girls. Especially item 12 seems to fit poorly, since it has a fit index of 19.1828 while the 5% critical value of a chi-square variable with six degrees of freedom is 12.6. This item may be subject to bias, but nothing conclusive can be said, since both groups are not yet on one scale. So next the item parameters were estimated using both boys and girls. The results of the evaluation of model fit are shown in Table 3. It can be seen that the model does

---

Insert Table 3 about here

---

not fit for boys and girls together. Candidates for bias are item 21, which has a fit index of 12.3204 for girls and 13.9846 for boys, and item 50, with fit indices of 11.9820 and 23.1616 for girls and boys, respectively. Inspection of the scaled deviates of item 21 shows that they are positive for girls and negative for boys. Since the scaled deviates are based on the difference of observed and expected frequencies, it can be concluded that the item favours the girls. In can be verified that this also holds for item 50. Item 12 is a different matter. It fits well for girls, but it does not fit for the boys. Inspection of the scaled deviates shows that the item discriminates too little for the second group. Using the terminology of Mellenbergh (1982,1983) items 21 and 50 show uniform bias, while item 12 is non-uniformly biased. However, the presence of biased items may still detract from the value of the sum score as a sufficient statistic for ability. Therefore in the next analysis the items 21 and 50 were considered to be different for boys and girls. For practical purposes, for the boys item 21 was labeled 121 and item 50 was labeled 150. In the parameter estimation item 21 and 121 and item 50 and 150 were treated as different items. In Table 4 the resulting

---

Insert Table 4 about here

---

evaluation of model fit is shown. It can be seen that the items 21, 121, 50 and 150 show a good fit. Further, the R_1 statistic drops from 299.1520 in the previous analysis to 250.9180, which is considerable given a drop of two degrees of freedom. However, global model fit is still not perfect. To investigate whether this could be contributed to item 12, several analyses were run. Splitting item 12 into different items for boys and girls resulted in an R_1 value of 245.3231 with 217 degrees of freedom, which still is poor. Removing the item from the boys' subscale did help ($R_1 = 243.159$, $df = 212$), but good fit was only achieved when the item was completely removed ($R_1 = 240.734$, $df = 207$). Summing up, evidence has been produced that items 21 and 50 are uniformly biased in favour of the girls, while item 12 is non-uniformly biased and discriminates too little for the boys. The second example concerns an examination in language comprehension of Dutch, which consisted of 20 polytomous items with 4 to 6 score points each. The procedure for detecting item bias is exactly the same as for dichotomous items, that is, first a number of Rasch- homogeneous subscales are identified using a procedure based on the R_1-statistic, then the differences in response probabilities between the two groups are evaluated for matched ability levels. Therefore the procedure will not be reiterated in detail, only the effects of the higher level of complexity of the data structure will be illustrated by showing some tables of counts on which the procedure is based. In Table 5 observed and expected frequencies for four items with four response categories are shown for the group of girls with scores from one to four. Notice that every item yields a distinct scaled deviate for every response category. The results of this score group and other score groups are summarized in Table 6. Again, for every item/category combination an index of fit is computed by squaring and summing the scaled deviates over subgroups. An index of item fit can be computed by summing the item/category indices over categories. At the bottom of the table it can be seen that the items make up a Rasch-scale. The scale was found by a process of elimination similar as the one described for dichotomous items. Introducing the boys resulted in $R_1 = 237.683$ with 177 degrees of freedom, so the scale did not fit both groups together. Splitting one of the items resulted in a significant improvement ($R_1 = 173.149$ with 161 degrees of freedom).

The point to be made from inspection of the tables is that interpreting item bias becomes rather complicated for polytomous items. The response categories of the

items are associated with different numbers of score points to be acquired. If an item favours one of the two groups, for this group some higher indexed categories will attract more responses that expected, while some lower indexed categories, but not necessarily the zero category, will attract fewer responses than expected. Especially if the number of response categories is large, detecting which group is favoured by the item may be complicated. So while searching for Rasch-homogeneous subscales can be carried out in much the same mechanical manner as for dichotomous items and biased items can be pinpointed in a statistically sound fashion, the interpretation of the results may, in most instances, be a tedious job.

References.

Andersen, E.B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 34,* 42 - 50.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Andrich, D. (19 78a). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38,* 665-680.

Andersen, E.B. (1980). *Discrete statistical models with social science applications.* Amsterdam: North-Holland Publishing Company.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory.* New York: Wiley.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. (1975). *Discrete multivariate analysis: theory and practice.* Cambridge (Mass.) : M.I.T. press.

Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B, 25,* 220-233.

Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika, 46,* 443-459.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests.* Wien: Verlag Hans Huber.

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

Fischer, G.H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3-26.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525-546.

Glas, C.A.W. (1989). *Contributions to estimating and testing Rasch models.* Thesis, University of Twente.

Glas, C.A.W. & Verhelst, N.D. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635-659.

Haberman, S.J. (1974). *The Analysis of Frequency Data.* Chicago: University of Chicago press.

Hambleton, R.K. & Rogers, H.J. (1989). Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2(4),* 313-334.

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

Kelderman, H, (1989). Item bias detection using loglinear IRT. *Psychometrika, 54,* 681-697.

Kok, F.G., Mellenbergh, G.J. & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22,* 295-303.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,*

Masters, G.N. & Wright, B.D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529 - 544.

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Mellenbergh, G.J. (1983). Conditional item bias methods. In S.H. Irvine & W.J.Berry (Eds.), *Human Assessment and Cultural Factors* (pp.293-302). New York: Plenum Press.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications,* (Second Edition). New York: Wiley.

Rasch, G. (1961). On the general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 321-333. Berkeley: University of California Press.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In M. Blegvad (Ed.), *Danish Yearbook of Philosophy*. Copenhagen: Munksgaard.

TABLE 1: EVALUATION OF MODEL FIT FOR THE GIRLS

OVERVIEW OF ITEM FIT

Item	Group-1	Group-2	Group-3	Group-4	Group-5	Group-6	Item Fit
1	.9539	-.4894	-.5201	.5075	-.7086	-.2740	2.2546
2	.4536	-1.1639	.5201	-.2266	-.3598	1.0431	3.0997
12	2.4841	-.6653	.8950	.0609	-.6228	-.9711	8.7490
15	.0093	-1.6490	-.3877	-.1332	1.7119	1.1783	7.2063
16	-1.7211	.0470	.6594	.7867	1.1162	.5821	5.6029
17	1.8075	-1.0339	-1.0674	.6589	.1703	-.6168	6.3188
19	.4132	-.7691	.5828	.0394	.3213	-.5801	1.5433
20	.2086	1.1698	-1.1354	-.0129	-.7168	.5668	3.5363
21	-.5188	.1046	-.4980	.5179	-.1273	1.2831	2.4588
22	-.5436	-1.2984	.3382	.4549	.9581	.7790	3.8274
23	-1.6872	.6186	1.1860	.5612	-.2028	-.3745	5.1322
24	.2516	.8687	.0289	-2.8093	.6866	.7670	9.7708
25	-.3619	-.0812	1.3287	-.6914	-.9440	.7985	3.9098
31	-.7423	1.2181	-.5695	-.2051	.0513	.1544	2.4276
33	-.8157	1.2858	-.0234	-.0003	-.2389	-.4710	2.5982
35	.6263	.6051	-.1067	-.8476	-.3516	-.3153	1.7113
38	-.3653	1.1492	.3123	.9173	-1.3554	-.6291	4.6260
44	.1139	-.7721	1.0760	-.5791	1.6335	-1.6881	7.6204
45	-.6578	-.2516	-.8836	1.5314	.9845	.1138	4.6043
46	.9270	-.1541	-1.1481	.4572	-.3744	.4308	2.7358
50	-.1925	-.1120	.6091	.5568	-.0843	-.8846	1.5202

OUTCOME OF THE R1-TEST:	86.6060
DF:	100
PROB(R1):	.8278

TABLE 2: EVALUATION OF MODEL FIT FOR THE BOYS

OVERVIEW OF ITEM FIT

Item	Group-1	Group-2	Group-3	Group-4	Group-5	Group-6	Item Fit
1	.9311	1.1718	-1.0582	-.6757	-1.4345	-.4748	6.0999
2	1.2253	-.2163	-.6966	.5540	-1.5503	-.6434	5.1577
12	3.3095	1.5571	-.4026	-.1881	-1.1810	-2.0527	19.1828
15	-.1473	-.0531	-.0994	-.5097	.7785	.5112	1.1616
16	-1.6850	-.0819	.6855	1.1445	.6832	1.1711	6.4638
17	-.2890	-.3729	.6726	-.7774	.4646	.5265	1.7724
19	-.3042	-.3734	-.3122	.9707	-.1183	.9669	2.2205
20	-.1854	1.0966	.5024	-1.3662	.5111	-1.1663	4.9776
21	-.4848	.0531	-.4130	.8175	.0751	1.0185	2.1197
22	-1.4908	-.9060	.7035	.5459	1.4769	.7352	6.5579
23	-.4798	-1.4734	-.0343	.6455	2.6518	.0430	9.8528
24	1.8194	.0015	.8486	-.6829	-1.9582	-.1241	8.3468
25	-.7278	.1847	-.7220	1.6328	.1192	1.1711	5.1369
31	-.0640	-1.0764	.0130	.4010	.8520	.9589	2.9690
33	-1.2530	-.5896	1.0272	1.4216	-.0815	-.1189	5.0145
35	.0019	-.2783	-.5291	-.2349	.6912	1.4809	3.0832
38	.9405	.8079	-1.0827	.7022	-.6131	-.5897	3.9259
44	-.9390	.0741	1.3067	-1.4776	1.3753	-.4002	6.8298
45	-1.3715	.5779	.7500	-.1956	-.2765	1.0438	3.9819
46	.5711	.3277	-.2117	-1.3302	-.4100	1.3701	4.2933
50	1.8432	-.0435	-1.3524	.5167	-.6828	-1.4142	7.9617

OUTCOME OF THE R1-TEST:	111.1115
DF:	100
PROB(R1):	.2104

TABLE 3: EVALUATION OF MODEL FIT FOR THE GIRLS AND BOYS TOGETHER

OVERVIEW OF ITEM FIT FOR GIRLS (CONTRIBUTION TO R1: 130.074)

Item	Group-1	Group-2	Group-3	Group-4	Group-5	Group-6	Item Fit
1	1.7244	.3207	.1733	.9576	-.1286	.1874	4.0752
2	.8133	-.7182	.7774	.0316	-.0930	1.1501	3.1141
12	2.1308	-.9490	.6203	-.1644	-.8566	-1.1895	8.0014
15	-1.2440	-2.9167	-1.4367	-.9622	1.0290	.5885	14.4497
16	-1.2700	.4569	.9806	1.0300	1.3350	.7978	6.2628
17	1.2749	-1.4665	-1.4028	.4417	-.0326	-.7850	6.5562
19	-.3156	-1.4598	.0209	-.4118	-.1045	-.9599	3.3330
20	-.3079	.5784	-1.6701	-.4891	-1.2380	.0809	4.9970
21	1.3815	1.7066	.9310	1.4762	.9266	1.8960	12.3204
22	-.3143	-.9100	.7080	.7756	1.2837	1.0846	4.8541
23	-1.3103	1.1072	1.6308	.9524	.2369	.0413	6.5671
24	.4815	1.3067	.5285	-2.2975	1.2638	1.3755	10.9859
25	-1.5182	-.9875	.7269	-1.2925	-1.5366	.4485	8.0415
31	-1.1738	.9086	-.8103	-.3729	-.0887	.0590	3.0104
33	-1.4433	.8154	-.4086	-.2876	-.5082	-.6887	3.7303
35	.2778	.3773	-.2622	-.9539	-.4288	-.3557	1.5086
38	-1.0188	.3880	-.3912	.3180	-2.0240	-1.2460	7.0918
44	.7009	-.1445	1.4858	-.1377	1.8853	-1.1954	7.7218
45	-.3103	.1705	-.4704	1.7818	1.2614	.3999	5.2724
46	1.0910	.1695	-.7843	.7681	-.0107	.7615	3.0041
50	1.6891	1.5451	1.8556	1.5211	.9745	.1878	11.9820

OVERVIEW OF ITEM FIT FOR BOYS (CONTRIBUTION TO R1: 169.078)

Item	Group-1	Group-2	Group-3	Group-4	Group-5	Group-6	Item Fit
1	.1429	.4578	-1.9754	-1.2312	-2.0260	-.8142	10.4154
2	.8910	-.6139	-1.1262	.3809	-1.8827	-.8319	6.8206
12	3.6536	1.8760	-.0693	.0365	-.9599	-1.8763	21.3167
15	1.2141	1.2560	1.1722	.2649	1.3425	.8752	7.0639
16	-2.1641	-.5236	.2779	.9465	.4875	1.0918	7.3602
17	.2304	.0789	1.0501	-.5472	.6255	.6193	2.2364
19	.4329	.3637	.4137	1.3528	.2438	1.1572	3.7193
20	.3286	1.7228	1.2150	-.8687	.9642	-.8032	6.8816
21	-2.2642	-1.6819	-2.2264	-.0793	-.8335	.6093	13.9846
22	-1.7728	-1.3105	.2373	.2375	1.2247	.5497	6.7750
23	-.8958	-2.0297	-.6485	.2720	2.4106	-.2140	11.2735
24	1.5139	-.4505	.1777	-1.2308	-2.5866	-.5827	11.0716
25	.4883	1.1351	.2262	1.9511	.5202	1.3125	7.3780
31	.3383	-.7239	.3048	.5454	.9579	1.0146	2.9757
33	-.6315	-.0432	1.4764	1.6382	.1391	.0211	5.2838
35	.3123	-.0311	-.3319	-.1432	.7499	1.5039	3.0533
38	1.6746	1.6249	-.1583	1.2835	-.0403	-.1853	7.1531
44	-1.5946	-.5355	.7876	-1.9713	1.1673	-.6572	9.1303
45	-1.7687	.1566	.3040	-.5043	-.5634	.9167	4.6574
46	.3514	-.0138	-.6677	-1.6840	-.7343	1.1973	5.3781
50	.2313	-1.8115	-3.3397	-.4491	-1.7507	-2.3252	23.1616

OUTCOME OF THE R1-TEST:	299.1520
DF:	220
PROB(R1):	.0003

TABLE 4: EVALUATION OF MODEL FIT FOR THE GIRLS AND BOYS TOGETHER

OVERVIEW OF ITEM FIT FOR GIRLS (CONTRIBUTION TO R1: 108.259)

Item	Group-1	Group-2	Group-3	Group-4	Group-5	Group-6	Item Fit
1	1.8640	.4024	.2268	.9867	-.0987	.2055	4.7137
2	.9140	-.6567	.8047	.0536	-.0747	1.1559	3.2585
12	2.3124	-.8057	.7565	-.0540	-.7431	-1.0827	8.2961
15	-1.0273	-2.7513	-1.3225	-.8836	1.0842	.6272	12.7232
16	-1.0947	.5389	1.0284	1.0593	1.3562	.8132	6.1691
17	1.4874	-1.3134	-1.2937	.5076	.0245	-.7424	6.4206
19	-.1007	-1.2921	.1403	-.3254	-.0320	-.9052	2.6256
20	-.1369	.7462	-1.5334	-.3781	-1.1287	.1670	4.3715
21	-.5665	.1081	-.4818	.5338	-.1036	1.2999	2.5502
22	-.0976	-.7524	.8125	.8462	1.3395	1.1223	5.0058
23	-1.0935	1.2486	1.7237	1.0173	.2944	.0810	6.8540
24	.6721	1.4807	.6687	-2.1862	1.3617	1.4483	11.8228
25	-1.3243	-.8765	.7879	-1.2387	-1.4910	.4700	7.1211
31	-.9574	1.0581	-.6969	-.2954	-.0250	.1020	2.6202
33	-1.2268	.9526	-.3078	-.2188	-.4497	-.6477	3.1768
35	.4950	.5304	-.1508	-.8722	-.3624	-.3100	1.5373
38	-.8390	.5599	-.2518	.4241	-1.9193	-1.1662	6.3045
44	.8582	-.0539	1.5312	-.0984	1.9041	-1.1690	8.0857
45	-.1116	.2869	-.3882	1.8220	1.2965	.4264	5.4280
46	1.3101	.3331	-.6607	.8493	.0619	.8084	3.6426
50	-.2397	-.1084	.6228	.5727	-.0606	-.8527	1.5158

OVERVIEW OF ITEM FIT FOR BOYS (CONTRIBUTION TO R1: 142.659)

Item	Group-1	Group-2	Group-3	Group-4	Group-5	Group-6	Item Fit
1	-.0065	.3647	-2.0668	-1.2743	-2.0648	-.8325	10.9853
2	.7994	-.6869	-1.1851	.3629	-1.9126	-.8459	7.0205
12	3.4685	1.7074	-.2384	-.0726	-1.0633	-1.9552	19.9615
15	1.0119	1.0971	1.0471	.2025	1.3028	.8534	5.7906
16	-2.3426	-.6297	.2073	.9207	.4663	1.0850	8.1696
17	.0290	-.0806	.9314	-.6115	.5844	.5979	1.9478
19	.2295	.1871	.2662	1.2865	.1885	1.1318	3.1300
20	.1631	1.5419	1.0371	-.9759	.8776	-.8630	5.9470
121	-.4476	.0555	-.4294	.8026	.0560	1.0075	2.0500
22	-1.9776	-1.4857	.1026	.1730	1.1828	.5252	7.8333
23	-1.1002	-2.2048	-.7826	.2117	2.3796	-.2411	12.4493
24	1.3273	-.6273	.0011	-1.3355	-2.6811	-.6356	11.5313
25	.3219	1.0305	.1419	1.9269	.4946	1.3043	6.8445
31	.1345	-.8964	.1700	.4827	.9139	.9926	2.9041
33	-.8344	-.2011	1.3634	1.5908	.0959	-.0032	5.1355
35	.1084	-.2011	-.4722	-.2113	.7041	1.4843	3.0186
38	1.4832	1.4398	-.3342	1.1895	-.1208	-.2339	5.8689
44	-1.7692	-.6446	.7196	-2.0221	1.1501	-.6752	9.9313
45	-1.9674	.0249	.2062	-.5549	-.6019	.9031	5.3996
46	.1484	-.1920	-.8219	-1.7692	-.7959	1.1718	5.8713
150	1.8787	-.0409	-1.3700	.5009	-.7048	-1.4369	8.2205

OUTCOME OF THE R1-TEST:	250.9180
DF:	218
PROB(R1):	.0624

TABLE 5: OBSERVED AND EXPECTED FREQUENCIES FOR GIRLS
WITH SCORES FROM 1 TO 4 (N=106)

ITEM	CAT	OBSERVED	EXPECTED	DEVIATE	SCALED DEVIATE
5	1	58	62.798	-4.798	-.801
	2	103	101.106	1.894	.271
	3	37	40.616	-3.616	-.684
	4	23	16.906	6.094	1.617
6	1	3	4.904	-1.904	-.869
	2	15	18.946	-3.946	-.953
	3	1	.677	.323	.394
	4	4	2.152	1.848	1.273
8	1	4	4.061	-.061	-.031
	2	11	9.954	1.046	.340
	3	0	1.141	-1.141	-1.072
	4	4	3.315	.685	.382
16	1	31	32.329	-1.329	-.250
	2	70	80.004	-10.004	-1.417
	3	12	8.601	3.399	1.191
	4	0	.324	-.324	-.570

Contribution to R1-statistic : 12.462

TABLE 6: EVALUATION OF MODEL FIT FOR THE GIRLS

OVERVIEW OF ITEM FIT

Item	Cat	Group-1	Group-2	Group-3	Group-4	Group-5	Group-6	Item Fit
5	1	-.8012	1.1094	-.3857	1.3026	.3918	-1.8118	7.1545
	2	.2711	-1.5841	-.7690	-1.5828	1.0782	1.0824	8.0136
	3	-.6836	.0620	.3207	.9624	-.8654	.3387	2.3638
	4	1.6167	.2311	.3062	-1.1487	-.3414	-.2196	4.2453
6	1	-.8686	-.7641	.6655	-1.8085	1.9890	.3840	9.1556
	2	-.9531	.0652	.8538	-.4862	-1.0679	1.3126	4.7414
	3	.3935	-.3295	-1.3894	-1.7324	1.7898	.1215	8.4134
	4	1.2732	1.1416	.4233	1.8760	-1.2270	-.7171	8.6425
8	1	-.0305	.0704	1.6339	-.3781	-.8889	-.1944	3.6464
	2	.3402	.5850	-.5103	-.8060	1.0640	-.7357	3.0412
	3	-1.0720	.4062	.9397	-1.8793	-.0891	.9194	6.5821
	4	.3822	-.9178	.3895	1.0137	.3428	-1.0993	3.4936
16	1	-.2505	.2576	2.1679	-.1985	-1.0951	.0076	6.0676
	2	-1.4170	.9864	-.1917	-.4398	.7858	.1700	3.8576
	3	1.1910	-1.4205	-2.4318	-.6313	.5917	1.2228	11.5938
	4	-.5703	1.0713	.2228	-.3024	.7482	-.5754	2.5050

OUTCOME OF THE R1-TEST:	88.5175
DF:	81
PROB(R1):	.2744

Recent Measurement and Research Department Reports:

91-1 N.D. Verhelst & N.H. Veldhuijzen. A New Algorithm For Computing Elementary Symmetric Functions And Their First And Second Derivatives.