DOCUMENT RESUME

ED 421 550 TM 028 876

AUTHOR Glas, Cees A. W.
TITLE Modification Indices for the 2PL and the Nominal Response Model. Research Report 98-04.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE 1998-00-00
NOTE 42p.
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Foreign Countries; *Goodness of Fit; Item Response Theory; Maximum Likelihood Statistics; *Test Items
IDENTIFIERS *Nominal Response Model; Rasch Model; Two Parameter Model

ABSTRACT

In this paper it is shown that various violations of the two parameter logistic (2PL) model can be evaluated using the Lagrange multiplier test (J. Aitchison and S. Silvey, 1958) or the equivalent difference score test. The tests focus on violation of local stochastic independence and insufficient capture of the form of the item characteristic curves. Primarily, the tests are item-oriented diagnostic tools, but taken together, they also serve the purpose of evaluation of global model fit. A useful feature of Lagrange multiplier statistics is that they are evaluated using maximum likelihood estimates of the null model only; that is, the parameters of alternative models need not be estimated. As numerical examples, an application on real data and some power studies are presented. (Contains 1 figure, 9 tables, and 33 references.) (Author/SLD)

Modification Indices for the 2PL and the Nominal Response Model

Research Report 98-04

Cees A.W. Glas

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Modification Indices for the 2-PL and the Nominal Response Model

Cees A.W. Glas

Abstract

In this paper, it is shown that various violations of the 2-PL model and the nominal response model can be evaluated using the Lagrange multiplier test or the equivalent efficient score test. The tests presented here focus on violation of local stochastic independence and insufficient capture of the form of the item characteristic curves. Primarily, the tests are item-oriented diagnostic tools, but taken together, they also serve the purpose of evaluation of global model fit. A useful feature of Lagrange multiplier statistics is that they are evaluated using maximum likelihood estimates of the null-model only, that is, the parameters of alternative models need not be estimated. As numerical examples, an application on real data and some power studies are presented.

Key words: efficient score test, item response theory, model fit, modification indices, 2-parameter logistic model, nominal response model, Lagrange multiplier test.

Introduction

Interestingly, evaluation of model fit has a long tradition in the Rasch model (Andersen, 1973, Martin Löf, 1973, 1974, Fischer, 1974, Kelderman, 1984, 1989, Molenaar, 1983, Glas, 1988, 1997, Glas & Verhelst, 1989, 1995) while contributions to the 2-PL model in this respect have been relatively few (Yen, 1981, Mislevy & Bock, 1990, Reiser, 1996, Glas, 1998). One of the reasons for this situation might be that the Rasch model and its variants have minimal sufficient statistics, which are very helpful for the derivation of the asymptotic distribution of the test statistics (see, for instance, Glas 1997). On the other hand, the 2-PL model is a more flexible model, so that the need for evaluation of model fit may be less stringent than in the case of the more restrictive Rasch model. However, also in the 2-PL model violations may occur which threaten the validity of the inferences made. In this paper, the focus will be on two violations: improper modeling of the form of the item characteristic curves (ICC's) and lack of local stochastic independence. In many respects, the model tests proposed here can be viewed as generalizations of two tests for the Rasch model: the R_1-test for evaluation of the assumption with respect to the form of the ICC's and the R_2-test for evaluation of the assumption of local independence (Glas, 1988, 1997, Glas & Verhelst, 1989, 1995).

The procedures proposed here are based on the Lagrange multiplier (LM) statistic (Aitchison & Silvey, 1958), rather than on likelihood ratio tests and Wald tests. This choice is made because LM tests only need ML estimates of the parameters of the model of the null-hypothesis. In the present case the null-model will be the 2-PL model, its generalization to polytomous data, the nominal response model (NRM, Bock, 1972) and a special case of the latter model, the generalized partial credit model (GPCM) by Muraki (1992). Generalization of the approach presented here to the 3-PL model is beyond the scope of the present paper and will be treated in a subsequent paper. In many instances, the parameters of the model of the alternative hypothesis will be quite complicated to estimate. But even if this is not the case, the procedure proposed here has advantages. In the sequel, hypothesis related to specific model violations will be tested for one item or pair of items at a time. If this was done using a Wald or likelihood ratio test, this would require computing new estimates for every test. So primarily, the procedures are meant as item-oriented diagnostic tools, However, below it will also be shown that the ensemble of the computed statistics can also serve the purpose of a global test of model fit.

Preliminaries

Consider items where the possible responses are be coded by the integers 0, 1, 2, 3, ..., m_i. Let item i have $m_i + 1$ response categories, indexed $g = 0, 1, ..., m_i$. Notice that dichotomous items are the special case where $m_i = 1$. The response of a person n to an item i will be represented by a vector $\boldsymbol{x}'_{ni} = (x_{ni0}, ..., x_{nig}, ..., x_{nim_i})$, where x_{nig} is a realization of the random variable X_{nig}; $x_{nig} = 1$ if the response is in category g and $x_{nig} = 0$ if this is not the case. The probability of scoring in category g of item i is given by

$$\psi_{ig}(\theta_n) = Pr(X_{nig} = 1 \mid \theta_n, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i) \propto \exp(\alpha_{ig}\theta_n - \beta_{ig}), \tag{1}$$

for $g = 0, 1, ..., m_i$, with the usual restriction $\alpha_{i0} = \beta_{i0} = 0$ to identify the model. Defining $\psi_{ig}(\theta_n)$ starting from $g = 0$ may seem a bit awkward here, but below it will prove very convenient. With the assumption of local independence between item responses, formulation (1) encompasses the 2-PL model for dichotomous items (Birnbaum, 1972), and the nominal response model (Bock, 1972, Thissen, 1991) for polytomous items. When the restriction $\alpha_{ig} = g\alpha_i$ is imposed, the model is the GPCM by Muraki (1992).

To introduce the LM tests, first some theory on MML estimation for IRT models must be summarized. The choice of an ability distribution is not essential to the theory presented here; it can either be the parametric (see Bock & Aitkin, 1982) or the non- parametric MML framework (see De Leeuw & Verhelst, 1986, Follmann, 1988). However, to make the presentation specific, the parametric framework will be assumed, and ability will be normally distributed with parameters μ and σ. Further, for reasons of simplicity, it will first be assumed that all respondents belong to the same population and have responded to the same set of items. Modern software for the 2-PL model, such as Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), also supports multiple populations and incomplete designs. Generalization of the methods to be presented to these specifications is straightforward and will be sketched in Section 7. Further, this software also supports Bayes modal estimation (Mislevy, 1986). This generalization will be discussed in Section 6.

Let $g(.; \mu, \sigma)$ be the density of θ_n. Since only one population is considered, the model can be identified introducing the restrictions $\mu = 0.0$ and $\sigma = 1.0$ and the remaining free parameters are the vectors of item parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The log-likelihood function of the parameters $\boldsymbol{\xi}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$ can be written as

$$\ln L(\boldsymbol{\xi}; \boldsymbol{X}) = \sum_n \ln Pr(\boldsymbol{x}_n; \boldsymbol{\xi}) \,, \tag{2}$$

where $\boldsymbol{x}_n$ is the response pattern of respondent n and $\boldsymbol{X}$ stands for the data matrix. To derive the MML estimation equations, it proves convenient to introduce the vector of derivatives

$$\boldsymbol{b}_n(\boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\xi}} \ln Pr(\boldsymbol{x}_n, \theta_n; \boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\xi}} [\ln Pr(\boldsymbol{x}_n \mid \theta_n, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \ln g(\theta_n; \mu, \sigma)] \tag{3}$$

with

$$Pr(\boldsymbol{x}_n \mid \theta_n, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_i \prod_{g=0}^{m_i} \psi_{ig}(\theta_n)^{x_{nig}} \,. \tag{4}$$

Adopting an identity by Louis (1982, also see, Glas, 1992), the first order derivatives of (2) with respect to $\boldsymbol{\xi}$ can be written as

$$\boldsymbol{h}(\boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\xi}} \ln L(\, \boldsymbol{\xi}; \boldsymbol{X}) = \sum_n E(\boldsymbol{b}_n(\boldsymbol{\xi}) \mid \boldsymbol{x}_n, \boldsymbol{\xi}) \,. \tag{5}$$

This identity greatly simplifies the derivation of the likelihood equations. For instance, it can be easily verified that the elements of $\boldsymbol{b}_n(\boldsymbol{\xi})$ are given by

$$b_n(\alpha_{ig}) = \theta_n(x_{nig} - \psi_{nig}) \tag{6}$$

and

$$b_n(\beta_{ig}) = \psi_{nig} - x_{nig}, \tag{7}$$

where ψ_{nig} is a short-hand notation for $\psi_{ig}(\theta_n)$. Combining these two expressions with (5), the likelihood equations for the item parameters are given by

$$\sum_n E(\theta_n x_{nig} \mid \boldsymbol{x}_n, \boldsymbol{\xi}) = \sum_n E(\theta_n \psi_{nig} \mid \boldsymbol{x}_n, \boldsymbol{\xi}) \tag{8}$$

and

$$\sum_n x_{nig} = \sum_n E(\psi_{nig} \mid \boldsymbol{x}_n, \boldsymbol{\xi}) . \tag{9}$$

For LM statistics, also the second order derivatives of the log-likelihood function are needed. It will prove convenient to define

$$\boldsymbol{H}(\boldsymbol{\xi}, \boldsymbol{\xi}) = -\frac{\partial^2 \ln L(\boldsymbol{\xi}; \boldsymbol{X})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} . \tag{10}$$

As with the derivation of the estimation equations, also for the derivation of the matrix of second order derivatives the theory by Louis (1982) can be used, and it follows that the observed information matrix, evaluated using MML estimates, is given by

$$\boldsymbol{H}(\boldsymbol{\xi}, \boldsymbol{\xi}) = -\sum_n [E(\boldsymbol{B}_n(\boldsymbol{\xi}, \boldsymbol{\xi}) \mid \boldsymbol{x}_n, \boldsymbol{\xi}) - E(\boldsymbol{b}_n(\boldsymbol{\xi}) \boldsymbol{b}_n(\boldsymbol{\xi})' \mid \boldsymbol{x}_n, \boldsymbol{\xi})], \tag{11}$$

where

$$\boldsymbol{B}_n(\boldsymbol{\xi}, \boldsymbol{\xi}) = \frac{\partial^2 \ln Pr(\boldsymbol{x}_n, \theta_n; \boldsymbol{\xi})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} . \tag{12}$$

Notice that the expressions for the second of the two right-hand terms of (11) can be directly derived from (6) and (7) , the expressions for evaluating $\boldsymbol{B}_n(\boldsymbol{\xi}, \boldsymbol{\xi})$ are found upon taking derivatives of these two expressions. The exact expressions for (11) can also be found in Glas (1998).

For the 2-PL, the NRM, and the GPCM, the exact expressions for the second order derivatives are still tractable, but for more complicated models, using (11) may become rather complicated. A solution to this problem may be using the Fischer information matrix,

$$\boldsymbol{H}(\boldsymbol{\xi}, \boldsymbol{\xi}) \approx \sum_n E(\boldsymbol{b}_n(\boldsymbol{\xi}) \mid \boldsymbol{x}_n, \boldsymbol{\xi}) E(\boldsymbol{b}_n(\boldsymbol{\xi}) \mid \boldsymbol{x}_n, \boldsymbol{\xi})', \tag{13}$$

(also, see Mislevy (1986)). Below it will be shown by numerical examples that this approximation proves satisfactory for computing LM tests.

Lagrange multiplier tests

The principle of the LM test (Aitchison & Silvey, 1958), and the equivalent efficient-score test (Rao, 1947) can be summarized as follows. Consider a null-hypothesis about a model with parameters $\boldsymbol{\phi}_0$. This model is a special case of a general model with parameters $\boldsymbol{\phi}$. In the present, case the special model is derived from the general model by fixing one or more parameters to known constants. Let $\boldsymbol{\phi}_0$ be partitioned as $\boldsymbol{\phi}_0' = (\boldsymbol{\phi}_{01}', \boldsymbol{\phi}_{02}') = (\boldsymbol{\phi}_{01}', \boldsymbol{c}')$, where $\boldsymbol{c}$ is the vector of the postulated constants. Let $\boldsymbol{h}(\boldsymbol{\phi})$ be the partial derivatives of the log-likelihood of the general model, so $\boldsymbol{h}(\boldsymbol{\phi}) = \partial \ln L(\boldsymbol{\phi}) / \partial \boldsymbol{\phi}$. This vector of partial derivatives gauges the change of the log-likelihood as a function of local changes in $\boldsymbol{\phi}$. Let $\boldsymbol{H}(\boldsymbol{\phi}, \boldsymbol{\phi})$ be defined as $-\partial^2 \ln L(\boldsymbol{\phi}) / \partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'$. Then the LM statistic is given by

$$LM = \boldsymbol{h}(\boldsymbol{\phi}_0)' \boldsymbol{H}(\boldsymbol{\phi}_0, \boldsymbol{\phi}_0)^{-1} \boldsymbol{h}(\boldsymbol{\phi}_0). \tag{14}$$

If this statistic is evaluated using the ML estimate of $\boldsymbol{\phi}_{01}$ and the postulated values of $\boldsymbol{c}$, it has an asymptotic χ^2-distribution with degrees of freedom equal to the number of parameters fixed (Aitchison & Silvey, 1958). An important computational aspect of the procedure is that at the point of the ML estimates $\hat{\boldsymbol{\phi}}_{01}$ the free parameters have a partial derivative equal to zero. Therefore, (14) can be computed as

$$LM(\boldsymbol{c}) = \boldsymbol{h}(\boldsymbol{c})' \boldsymbol{W}^{-1} \boldsymbol{h}(\boldsymbol{c}) \tag{15}$$

with

$$\boldsymbol{W} = \boldsymbol{H}_{22}(\boldsymbol{c}, \boldsymbol{c}) - \boldsymbol{H}_{21}(\boldsymbol{c}, \hat{\boldsymbol{\phi}}_{01}) \boldsymbol{H}_{11}(\hat{\boldsymbol{\phi}}_{01}, \hat{\boldsymbol{\phi}}_{01})^{-1} \boldsymbol{H}_{12}(\hat{\boldsymbol{\phi}}_{01}, \boldsymbol{c}),$$

where the partitioning of $\boldsymbol{H}(\phi_0, \phi_0)$ into $\boldsymbol{H}_{22}(\boldsymbol{c}, \boldsymbol{c})$, $\boldsymbol{H}_{21}(\boldsymbol{c}, \hat{\phi}_{01})$, $\boldsymbol{H}_{11}(\hat{\phi}_{01}, \hat{\phi}_{01})$, and $\boldsymbol{H}_{12}(\hat{\phi}_{01}, \boldsymbol{c})$ is according to the partition $\phi_0' = (\phi_{01}', \phi_{02}') = (\phi_{01}', \boldsymbol{c}')$.

Notice that $\boldsymbol{H}(\phi_{01}, \phi_{01})$ also plays a role in the Newton-Raphson procedure for solving the estimation equations and in computation of the observed information matrix. So its inverse will usually by available at the end of the estimation procedure. Further, if the validity of the model of the null-hypothesis is tested against various alternative models, the computational work is reduced by the fact that the inverse of $\boldsymbol{H}(\hat{\phi}_{01}, \hat{\phi}_{01})$ is already available and the order of $\boldsymbol{W}$ is equal to the number of parameters fixed. It is advisable to keep the number of fixed parameters small to keep the interpretation of the outcome of the test tractable. This interpretation is supported by observing that the value of (15) depends on the magnitude of $\boldsymbol{h}(\boldsymbol{c})$, that is, on the first order derivatives with respect to the parameters ϕ_{02} evaluated in $\boldsymbol{c}$. If the absolute values of these derivatives are large, the fixed parameters are bound to change once they are set free, and the test is significant, that is, the special model is rejected. If the absolute values of these derivatives are small, the fixed parameters will probably show little change should they be set free, that is, the values at which these parameters are fixed in the special model are adequate and the test is not significant.

Besides a test of significance, this approach also provides information with respect to the direction in which the fixed parameters will change when set free. This is done by computing a new value of the fixed parameters, say ϕ_{02}^*, by performing one Newton-Raphson step, that is,

$$\phi_{02}^* = \boldsymbol{c} + \boldsymbol{W}^{-1}\boldsymbol{h}(\boldsymbol{c}) \,. \tag{16}$$

Below, this new value ϕ_{02}^*, will be called a modification index. The covariance matrix of ϕ_{02}^* can be approximated by $\boldsymbol{W}$. Assuming asymptotic normality of the estimates, it can then be tested whether ϕ_{02}^* significantly differs from $\boldsymbol{c}$, which boils down to performing the Rao (1947) efficient score test.

Evaluation of the Fit of Item Characteristic Curves

For dichotomous items, Lord (1980, pp.46-49) has pointed out that the expected number right score $\sum_i \psi_{i1}(\theta)$ and ability θ are the same things expressed on different scales of measurement. The important difference is that the measurement scale of the expected number right score depends on the test, while the measurement scale of θ is independent of the items

in the test. For polytomous items, the situation is more complicated, in fact, Hemker, Sijtsma, Molenaar and Junker (1996) have shown that the unweighted sum score does not necessarily have a monotone likelihood ratio in θ. However, usually the unweighted sum score and the associated estimate of θ will highly correlate.

The idea of the LM test and modification index presented here will be to partition the latent ability continuum into a number of segments, and to evaluate whether an item's ICC conforms the form predicted by the null-model in each of these segments. However, to be able to properly define an LM statistic, the actual partitioning will take place on the observed total score scale rather than on the θ scale. As already mentioned above, the LM tests and modification indices developed here focus on specific items. So let the item of interest be labeled i, while the other items are labeled $j = 1, 2, ..., i-1, i+1, ..., K$. Let $\boldsymbol{x}_n^{(i)}$ be a response pattern without item i, and let $r(\boldsymbol{x}_n^{(i)})$ be the unweighted sum score on this response pattern, that is,

$$r(\boldsymbol{x}_n^{(i)}) = \sum_{j \neq i} \sum_g g x_{njg} \,. \tag{17}$$

The possible scores $r(\boldsymbol{x}_n^{(i)})$ will be partitioned into S_i disjoint subsets; the index i signifies that this partition may be different for every item i. Consider the ordered boundary scores $r_0 < r_1 < r_2, ..., < r_s <, ..., < r_{S_i}$, with $r_0 = 0$ and $r_{S_i} = \sum_{j \neq i} m_j$. Further, define

$$w(s, \boldsymbol{x}_n^{(i)}) = \begin{cases} 1 & \text{if } r_{s-1} <= r(\boldsymbol{x}_n^{(i)}) < r_s, \\ 0 & \text{otherwise}, \end{cases} \tag{18}$$

so $w(s, \boldsymbol{x}_n^{(i)})$ is an indicator function which assumes a value equal to one if the unweighted sum score of response pattern $\boldsymbol{x}_n^{(i)}$ is in score range s. Because a partition of the score range also induces a partition of the sample of respondents, the term sub-sample will be used to signify groups of respondents with a sum score in a certain subset of the score range. The choice of the number of subsets S_i and the choice of the boundary scores will be returned to below.

First, the case of the 2-PL and the NRM will be considered, generalization to the GPCM will be sketched at the end of this section. The essence of the approach is introducing an alternative model with discrimination parameters $\alpha_{ig} + \gamma_{igs}$ and $\beta_{ig} + \delta_{igs}$. Consider a model where the probability of scoring in category g of item i, conditional on $w(s, \boldsymbol{x}_n^{(i)})$, is given by

$$\eta_{igs}(\theta_n) = Pr(X_{nig} = 1 \mid w(s, \boldsymbol{x}_n^{(i)}) = 1, \theta_n, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\delta}_i)$$

$$\propto \exp((\alpha_{ig} + \gamma_{igs})\theta_n - (\beta_{ig} + \delta_{igs})), \tag{19}$$

for $g = 0, 1, ..., m_i$. Under the null-model, that is, the 2-PL model or NRM, γ_{igs} and δ_{igs} will be equal to zero. In the alternative model, γ_{igs} and δ_{igs} are free parameters, which gauge the deviation of the discrimination and difficulty parameters in the sub-groups from the values α_{ig} and β_{ig}. Some restrictions need to be imposed to identify this model. For instance, the restrictions $\gamma_{i0s} = \delta_{i0s} = 0$ are imposed in addition to the usual restrictions $\alpha_{i0} = \beta_{i0} = 0$ to identify (19) for fixed s. Further, the complete set of S_i probabilities (19) can be identified using the restrictions $\gamma_{igS_i} = \delta_{igS_i} = 0$. Under this parametrization, α_{ig} and β_{ig} are the discrimination and difficulty parameters of item i in subgroup S_i and γ_{igs} and δ_{igs}, $s = 1, ..., S_i - 1$ are the deviations from this baseline in the other subgroups. An alternative to this parametrization will be considered in the section where a numerical example will be given.

The probability of a response pattern $\boldsymbol{x}_n$ is given by

$$Pr(\boldsymbol{x}_n \mid \theta_n, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}_i, \boldsymbol{\delta}_i) =$$

$$Pr(\boldsymbol{x}_{ni} \mid w(s, \boldsymbol{x}_n^{(i)}), \theta_n, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\delta}_i) Pr(\boldsymbol{x}_n^{(i)} \mid \theta_n, \boldsymbol{\alpha}, \boldsymbol{\beta}) =$$

$$\prod_{g=0}^{m_i} \prod_s \eta_{igs}(\theta_n)^{x_{ig} w(s, \mathbf{x}_n^{(i)})} \prod_{j \neq i} \prod_{h=0}^{m_j} \psi_{jh}(\theta_n)^{x_{jh}}, \tag{20}$$

where $\boldsymbol{x}_{ni}$ stands for the response on item i. Derivation of the LM test proceeds as follows. Let $\eta_{igs}(\theta_n)$ be abbreviated η_{nigs}. For respondents n with a sum score in category s, that is, $w(s, \boldsymbol{x}_n^{(i)}) = 1$, it holds that

$$b_n(\gamma_{igs}) = \theta_n(x_{nig} - \eta_{nigs}) \tag{21}$$

and

$$b_n(\delta_{igs}) = \eta_{nigs} - x_{nig}, \tag{22}$$

so the elements of the vectors of first order derivatives $\boldsymbol{h}(\boldsymbol{\gamma}_i)$ and $\boldsymbol{h}(\boldsymbol{\delta}_i)$ are given by

$$\sum_{n|w(s,\mathbf{x}_n^{(i)})=1} E(b_n(\gamma_{igs}) \mid \boldsymbol{x}_n,\ \boldsymbol{\xi}, \boldsymbol{\gamma}_i, \boldsymbol{\delta}_i) \tag{23}$$

and

$$\sum_{n|w(s,\mathbf{x}_n^{(i)})=1} E(b_n(\delta_{igs}) \mid \boldsymbol{x}_n,\ \boldsymbol{\xi}, \boldsymbol{\gamma}_i, \boldsymbol{\delta}_i). \tag{24}$$

Notice that from inspection of (22) and (24), it follows that the $h(\delta_{igs})$ is the difference of the observed number of persons of sub-sample s scoring in category g of item i, and its expected value. A the test for the simultaneous hypothesis $\gamma_{igs} = 0$ and $\delta_{igs} = 0$, for $s = 1, ..., S_i - 1$ and $g = 1, ..., m_i$, can be based on a statistic $LM(\boldsymbol{\gamma}_i, \boldsymbol{\delta}_i)$, which is defined by (15) with $\boldsymbol{\phi}'_{02} = \boldsymbol{c}' = (\boldsymbol{\gamma}'_i, \boldsymbol{\delta}'_i)$. When $LM(\boldsymbol{\gamma}_i, \boldsymbol{\delta}_i)$ is evaluated using MML estimates of the null-model, that is, with MML estimates of $\boldsymbol{\xi}$ and with $\boldsymbol{\gamma}_i = \mathbf{0}$, and $\boldsymbol{\delta}_i = \mathbf{0}$, $LM(\boldsymbol{\gamma}_i, \boldsymbol{\delta}_i)$ has a asymptotic χ^2-distribution with $2m_i(S_i - 1)$ degrees of freedom. It is also possible to define separate tests for the hypothesis $\gamma_{igs} = 0$, for $s = 1, ..., S_i - 1$ and $g = 1, ..., m_i$, and the hypothesis $\delta_{igs} = 0$, for $s = 1, ..., S_i - 1$ and $g = 1, ..., m_i$. The first test, say $LM(\boldsymbol{\gamma}_i)$, can be based on the first order derivatives $\boldsymbol{h}(\boldsymbol{\gamma}_i)$. This statistic $LM(\boldsymbol{\gamma}_i)$ has an asymptotic χ^2-distribution with $m_i(S_i - 1)$ degrees of freedom. In the same manner, a test based on a statistic $LM(\boldsymbol{\gamma}_i)$ can be defined for the hypothesis $\delta_{igs} = 0$, for $s = 1, ..., S_i - 1$ and $g = 1, ..., m_i$, which also has an asymptotic distribution with $m_i(S_i - 1)$ degrees of freedom.

Insert Table 1 about here

The exact expressions for the matrices of second order derivatives needed for evaluating (15) in the present case are found as follows. In the previous section, it was shown that the observed information matrix for the null-model, $\boldsymbol{H}_{11}(\boldsymbol{\phi}_{01}, \boldsymbol{\phi}_{01})$, with $\boldsymbol{\phi}_{01} = \boldsymbol{\xi}$ can

be derived using (11). This identity can also be used for deriving $\boldsymbol{H}_{22}(\boldsymbol{c},\boldsymbol{c})$ and $\boldsymbol{H}_{21}(\boldsymbol{c},\boldsymbol{\phi}_{01})$, with $\boldsymbol{c}' = (\boldsymbol{\gamma}_i', \boldsymbol{\delta}_i')$. In Table 1, expressions are given for $B_n(\phi_a, \phi_b)$, where ϕ_a is equal to γ_{igs}, γ_{ihs}, δ_{igs}, or δ_{ihs}, and ϕ_b is equal to γ_{igs}, γ_{ihs}, δ_{igs}, δ_{ihs}, α_{ig}, α_{ih}, β_{ig}, or β_{ih}. It is easily verified that elements $B_n(\gamma_{igs}, \gamma_{iht})$, $B_n(\delta_{igs}, \gamma_{iht})$ and $B_n(\delta_{igs}, \delta_{iht})$ are equal to zero if $s \neq t$. Further, if $i \neq j$, $B_n(\gamma_{igs}, \gamma_{jhs}) = 0$, $B_n(\delta_{igs}, \delta_{jhs}) = 0$ and $B_n(\gamma_{igs}, \delta_{jhs}) = 0$. The expression for $\boldsymbol{H}_{22}(\boldsymbol{c},\boldsymbol{c})$ can now be derived applying (11). For instance, the elements $E(B_n(\gamma_{igs}, \gamma_{ihs}) \mid \boldsymbol{x}_n, \boldsymbol{\xi}, \boldsymbol{\gamma}_i, \boldsymbol{\delta}_i)$ and $E(b_n(\gamma_{igs})b_n(\gamma_{ihs}) \mid \boldsymbol{x}_n, \boldsymbol{\xi}, \boldsymbol{\gamma}_i, \boldsymbol{\delta}_i)$ must be summed over all respondents with $w(s, \boldsymbol{x}_n) = 1$. The expressions for $\boldsymbol{H}_{21}(\boldsymbol{c}, \boldsymbol{\phi}_{01})$ are computed in a similar manner.

Similar tests can also be derived for the GPCM (Muraki, 1992). In this model, every item i has but one discrimination parameter α_i. Therefore, the shape of the ICC's are evaluated introducing $\alpha_i + \gamma_{is}$ and $\beta_{ig} + \delta_{igs}$ and testing the null- hypotheses $\gamma_{is} = 0$, $\delta_{igs} = 0$, or both $\gamma_{is} = 0$ and $\delta_{igs} = 0$, for $s = 1, ..., S_i - 1$ and $g = 1, ..., m_i$. Again, these tests can be based on statistics $LM(\boldsymbol{\gamma}_i)$, $LM(\boldsymbol{\delta}_i)$ and $LM(\boldsymbol{\gamma}_i, \boldsymbol{\delta}_i)$, which have $S_i - 1$, $m_i(S_i - 1)$ and $(m_i + 1)(S_i - 1)$ degrees of freedom, respectively. Since the GPCM is derived from the NRM by introducing the linear restrictions $\alpha_{ig} = g\alpha_i$, the matrix $\boldsymbol{H}_{11}(\boldsymbol{\phi}_{01}, \boldsymbol{\phi}_{01})$ for the GPCM can derived from the equivalent matrix for the NRM by pre- and post- multiplying the latter with the matrix of these linear restrictions and its transpose, respectively.

The definitions of $\boldsymbol{H}_{22}(\boldsymbol{c},\boldsymbol{c})$ and $\boldsymbol{H}_{21}(\boldsymbol{c}, \boldsymbol{\phi}_{01})$ are changed accordingly.

Evaluation of Local Stochastic Independence

Evaluation of local independence will be based on alternative models which are generalizations of models proposed by Kelderman (1984) and Jannarone (1986) in the framework of the Rasch model. To grasp the flavor of these models, they will be presented here for dichotomous items first. Let item i and item j be two items where the responses are dependent. Consider a model given by

$$Pr(\, x_i, x_j \mid \theta, \alpha_i, \alpha_j, \beta_i, \beta_j, \gamma_{ij}, \delta_{ij}) \propto$$

$$\exp[x_i(\alpha_i\theta - \beta_i) + x_j(\alpha_j\theta - \beta_j) + x_ix_j(\gamma_{ij}\theta - \delta_{ij})] \,, \tag{25}$$

where x_i and x_j take the values 0 or 1. In Table 2, the probabilities of the combinations of

x_i and x_j are cross-tabulated. From inspection of this table, it can be seen that γ_{ij} and δ_{ij} are parameters modeling the association between the two items. First consider a model without γ_{ij}, which is the 2-PL model version of the Kelderman (1984) model.

Insert Table 2 about here

In this model δ_{ij} represents the addition the item difficulty parameters β_i and β_j to account for the probability of a simultaneous correct response to the two items. In the generalization of the Jannarone (1986) model to the 2-PL model, besides an additional location parameter δ_{ij}, also an additional discrimination parameter γ_{ij} is added. This parameter accounts for interaction between the probability of a simultaneous correct response to the two items and the ability dimension θ.

This approach to modeling dependence between item responses can be generalized further to polytomous items by adding the appropriate number of rows and columns to the cross-tabulation of Table 2 and adding the parameters needed to model the additional row and column effects. As a result, the model for a simultaneous response to item i and item j becomes

$$\begin{aligned} \zeta_{nigjh} &= Pr(X_{nig} = 1, X_{njh} = 1 \mid \theta_n, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_{ij}, \boldsymbol{\delta}_{ij}) \\ &\propto \exp[(\alpha_{ig}\theta_n - \beta_{ig}) + (\alpha_{jh}\theta_n - \beta_{jh}) + (\gamma_{igjh}\theta_n - \delta_{igjh})], \end{aligned} \tag{26}$$

where $\gamma_{igjh} = \delta_{igjh} = 0$, if either $g = 0$ or $h = 0$. The probability of a response pattern changes from (4) to

$$Pr(\boldsymbol{x}_n \mid \theta_n, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}_{ij}, \boldsymbol{\delta}_{ij}) = \prod_{g=0}^{m_i} \prod_{h=0}^{m_j} \zeta_{nigjh}^{x_{ig}x_{jh}} \prod_{l \neq i,j} \prod_{k=0}^{m_k} \psi_{nlk}^{x_{lk}}. \tag{27}$$

LM tests and modification indices for assessing lack of local dependence can be based on derivatives of the log-likelihood with respect to γ_{igjh} and δ_{igjh}, evaluated under the null-model where $\gamma_{igjh} = 0$ and $\delta_{igjh} = 0$. Finding these derivatives, denoted $\boldsymbol{h}(\boldsymbol{\gamma}_{ij})$ and $\boldsymbol{h}(\boldsymbol{\delta}_{ij})$, again proceeds using expression (5). So inserting

$$b_n(\gamma_{igjh}) = \theta_n(x_{nig}x_{njh} - \zeta_{nigjh}) \tag{28}$$

and

$$b_n(\delta_{igjh}) = \zeta_{nigjh} - x_{nig}x_{njh} \tag{29}$$

into (5) produces the desired expressions. Notice that the $h(\delta_{igjh})$ is the difference between observing simultaneous responses $\sum_n x_{nig}x_{njh}$ and its expected value $\sum_n E(\zeta_{nigjh} \mid \boldsymbol{x}_n, \boldsymbol{\xi}, \boldsymbol{\gamma}_{ij}, \boldsymbol{\delta}_{ij})$. In the same manner, the expression for $h(\gamma_{igjh})$ is the difference between $\sum_n x_{nig}x_{nih}E(\theta_n \mid \boldsymbol{x}_n, \boldsymbol{\xi}, \boldsymbol{\gamma}_{ij}, \boldsymbol{\delta}_{ij})$ and $\sum_n E(\zeta_{nigjh}\theta_n \mid \boldsymbol{x}_n, \boldsymbol{\xi}, \boldsymbol{\gamma}_{ij}, \boldsymbol{\delta}_{ij})$.

A test for the composite null-hypothesis $\delta_{igjh} = 0$, and $\gamma_{igjh} = 0$, for $g = 1, ..., m_i$ and $h = 1, ..., m_j$, can be based on $LM(\boldsymbol{\gamma}_{ij}, \boldsymbol{\delta}_{ij})$, which is defined by (15) with $\boldsymbol{\phi}'_{02} = \boldsymbol{c}' = (\boldsymbol{\gamma}'_{ij}, \boldsymbol{\delta}'_{ij})$. When this statistic is evaluated using MML estimates of the null-model, it has an asymptotic χ^2- distribution with $2m_im_j$ degrees of freedom.

The matrix of weights $\boldsymbol{W}$ defined in (15), can again be found using (11). Therefore, expressions for $B_n(\phi_a, \phi_b)$ are needed, where ϕ_a and ϕ_b are γ_{igjh} and δ_{igjh} or a parameter of the null-model. The needed expressions are tabulated in Table 3, they easily follow from taking derivatives of (28) and (29).

Insert Table 3 about here

As in the previous section, also here special tests can defined for the hypothesis $\gamma_{igjh} = 0, g = 1, ..., m_i$ and $h = 1, ..., m_j$, and $\delta_{igjh} = 0, g = 1, ..., m_i$ and $h = 1, ..., m_j$. These tests, denoted $LM(\boldsymbol{\gamma}_{ij})$ and $LM(\boldsymbol{\delta}_{ij})$ are defined by (15) with $\boldsymbol{c} = \boldsymbol{\gamma}_{ij}$ and $\boldsymbol{c} = \boldsymbol{\delta}_{ij}$, respectively. They both have m_im_j degrees of freedom. Analogous to the previous section, tests for the GPCM can be defined as special cases of tests for the NRM. These tests, demoted $LM(\boldsymbol{\gamma}_{ij})$, $LM(\boldsymbol{\delta}_{ij})$ and $LM(\boldsymbol{\gamma}_{ij}, \boldsymbol{\delta}_{ij})$ have one, m_im_j and $2m_im_j$ degrees of freedom, respectively.

Modification Indices in a Bayes Modal Framework

It is well-known that item parameter estimates in the 2-PL model (and the 3-PL model,

which is beyond the scope of the present paper) are sometimes hard to obtain, because the parameters are poorly determined by the available data, in the sense that in the region of the ability scale where the respondents are located, the ICC's can be appropriately described by a large number of sets of item parameter values. To obtain "reasonable" and finite estimates, Mislevy (1986) considers a number of Bayesian approaches, entailing the introduction of prior distributions on the parameters. In the present section, it will be shown how the LM tests and modification indices presented above can accommodate these assumptions. In particular, two approaches will be studied, in the first approach the prior distribution is fixed, in the second approach, often labeled an empirical Bayes approach, the parameters of the prior distribution are estimated along with the other parameters. Let $p(\xi \mid \eta)$ be the prior density of the ξ, $\xi' = (\alpha', \beta')$, characterized by parameters η, which in turn follow a density $p(\eta)$. In a Bayes model framework, parameters estimates are computed by maximizing the posterior density of ξ, which is proportional to $\ln L(\xi; X) + \ln p(\xi \mid \eta) + \ln p(\eta)$.

First, the prior distribution of ξ will be considered known. Let $d(\xi) = \partial \ln p(\xi \mid \eta)/\partial \xi$ and $D(\xi, \xi) = -\partial^2 \ln p(\xi \mid \eta)/\partial \xi \partial \xi'$. The first order derivatives of the posterior with respect to ξ, say $h^*(\xi)$, are given by $h^*(\xi) = h(\xi) + d(\xi)$, where $h(\xi)$ is defined in (5), and the Bayes modal estimates are found upon solving $h^*(\xi) = 0$. The opposite of the second order derivatives of the posterior with respect to ξ, say $H^*(\xi, \xi)$, are given by $H^*(\xi, \xi) = H(\xi, \xi) + D(\xi, \xi)$, where $H(\xi, \xi)$ is defined in (10). Substituting $H^*(\xi, \xi)$ for $H_{11}(\xi, \xi)$ in the above LM statistics defines the comparable statistics for a Bayes modal framework with a fixed prior.

In an empirical Bayes framework, the parameters η, are estimated. Consider the definitions of Section 3. The parameter vector ϕ_0 was partitioned (ϕ_{01}', ϕ_{02}'). In the present context, ϕ_{01} is the concatenation (ξ', η') and ϕ_{02} can be γ_i , δ_i , or their concatenation, or γ_{ij} , δ_{ij} or their concatenation, all depending on the hypothesis considered. The first order derivatives of the posterior distribution are given by $h^*(\phi_0) = \partial \ln L(\xi; X)/\partial \phi_0 + \partial \ln p(\xi \mid \eta)/\partial \phi_0 + \partial \ln p(\eta)/\partial \phi_0$, which will be written as

$h^*(\phi_0) = h(\phi_0) + d(\phi_0) + g(\phi_0)$. So empirical Bayes modal estimation entails solving $h^*(\xi) = 0$ and $h^*(\eta) = 0$, that is, $h^*(\xi) = h(\xi) + d(\xi) = 0$ and $h^*(\eta) = d(\eta) + g(\eta) = 0$.

The opposite matrix of second order derivatives will be partitioned

$$H^*(\phi_0, \phi_0) = \begin{pmatrix} H^*_{11}(\xi, \xi) & H^*_{11}(\xi, \eta) & H^*_{12}(\xi, \phi_{02}) \\ H^*_{11}(\eta, \xi) & H^*_{11}(\eta, \eta) & H^*_{12}(\eta, \phi_{02}) \\ H^*_{21}(\phi_{02}, \xi) & H^*_{21}(\phi_{02}, \eta) & H^*_{22}(\phi_{02}, \phi_{02}) \end{pmatrix}.$$

Let $\boldsymbol{H}(\phi_0, \phi_0) = -\partial^2 \ln L(\boldsymbol{\xi}; \boldsymbol{X})/\partial\phi_0\partial\phi_0'$, $\boldsymbol{D}(\phi_0, \phi_0) = -\partial^2 \ln p(\boldsymbol{\xi} \mid \boldsymbol{\eta})/\partial\phi_0\,\partial\phi_0'$, and $\boldsymbol{G}(\phi_0, \phi_0) = -\partial^2 \ln p(\boldsymbol{\eta})/\partial\phi_0\partial\phi_0'$. Then the opposite matrix of second order derivatives becomes

$$\boldsymbol{H}^*(\phi_0, \phi_0) = \begin{pmatrix} \boldsymbol{H}_{11}(\boldsymbol{\xi}, \boldsymbol{\xi}) + \boldsymbol{D}(\boldsymbol{\xi}, \boldsymbol{\xi}) & \boldsymbol{D}(\boldsymbol{\xi}, \boldsymbol{\eta}) & \boldsymbol{H}^*_{12}(\boldsymbol{\xi}, \phi_{02}) \\ \boldsymbol{D}(\boldsymbol{\eta}, \boldsymbol{\xi}) & \boldsymbol{D}(\boldsymbol{\eta}, \boldsymbol{\eta}) + \boldsymbol{G}(\boldsymbol{\eta}, \boldsymbol{\eta}) & 0 \\ \boldsymbol{H}^*_{21}(\phi_{02}, \boldsymbol{\xi}) & 0 & \boldsymbol{H}^*_{22}(\phi_{02}, \phi_{02}) \end{pmatrix}.$$

Replacing $\boldsymbol{H}(\phi_0, \phi_0)$ in the above statistics by $\boldsymbol{H}^*(\phi_0, \phi_0)$ gives the definitions the equivalent statistics for the empirical Bayes modal framework. In the numerical examples given below, for dichotomous items, a fixed normal prior on the logarithm of α_i with parameters $\mu_{\ln\alpha}, \sigma_{\ln\alpha}$ will be used. For the empirical Bayes example, the natural conjugate prior for the normal distribution will be used, which is normal for $\mu_{\ln\alpha}$ given $\sigma_{\ln\alpha}$ and inverted Wishart for $\sigma_{\ln\alpha}$ (Ando & Kaufman, 1965). For details on this procedure one is referred to Mislevy (1986).

Multiple Populations and Incomplete Designs

Above, for reasons of simplicity, it was assumed that all respondents were drawn from the same population and responded to the same set of items. Generalization to a situation where this is not the case proceeds as follows. Firstly, it will be assumed that Q populations have normal ability distributions indexed by μ_q and σ_q, $q = 1, ..., Q$. Further, $q(n)$ is the population to which respondent n belongs. To identify the model, the first ability distribution will be fixed to standard normal, and the definition of the vector of free model parameters $\boldsymbol{\xi}$ is now extended to $\boldsymbol{\xi}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \mu_2, \sigma_2, ..., \mu_Q, \sigma_Q)$. Secondly, a missing data indicator $\boldsymbol{z}_n$ will be introduced. This vector has elements z_{ni} equal to one if a response of person n to item i is observed, and zero otherwise. In the present context, it will be assumed that the ignorability principle by Rubin (1976) holds, that is, the missing data indicator does not depend on the unobserved responses. As a consequence, parameters can be estimated using a likelihood function or a posterior distribution that is conditional on the value of the missing data indicator. Therefore, (4) and (5) now become

$$\begin{aligned} \boldsymbol{b}_n(\boldsymbol{\xi}) &= \frac{\partial}{\partial\boldsymbol{\xi}}[\ln Pr(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \theta_n, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \ln g(\theta_n; \mu_{q(n)}, \sigma_{q(n)})] \\ &= \frac{\partial}{\partial\boldsymbol{\xi}}[\sum_i \sum_g x_{nig} z_{nig} \ln \psi_{nig} + \ln g(\theta_n; \mu_{q(n)}, \sigma_{q(n)})]\,. \end{aligned} \tag{30}$$

The likelihood equations for the population parameters are derived upon observing that

$$b_n(\mu_{q(n)}) = (\theta_n - \mu_{q(n)})\sigma^{-2}_{q(n)} \tag{31}$$

and

$$b_n(\sigma_{q(n)}) = -\sigma^{-1}_{q(n)} + (\theta_n - \mu_{q(n)})^2\sigma^{-3}_{q(n)}, \tag{32}$$

Again using (5), first order derivatives can be derived, and estimation equations are given by

$$\mu_q = \frac{1}{N_q} \sum_{n|n(q)=q} E(\theta_n \mid \boldsymbol{x}_n, \boldsymbol{z}_n, \boldsymbol{\xi}) \tag{33}$$

and

$$\sigma^2_q = \frac{1}{N_q} \sum_{n|n(q)=q} E(\theta^2_n \mid \boldsymbol{x}_n, \boldsymbol{z}_n, \boldsymbol{\xi}) - \mu^2_q, \tag{34}$$

where N_q is the number of respondents in the sample of population q. First order derivatives for item parameters are derived from (8) and (9) by replacing the summations in these equations by summations over respondents n with z_{ni} equal to one, that is, the estimation equation for the parameters of item i only depends persons who have actually responded to item i. In the same manner, expressions for first order derivatives can be derived for item parameters of alternative models and for second order derivatives of item parameters.

Imputing these generalized definitions of $\boldsymbol{h}(\boldsymbol{\phi})$ and $\boldsymbol{H}(\boldsymbol{\phi}, \boldsymbol{\phi})$ into the definitions of the tests for local independence, $LM(\boldsymbol{\gamma}_{ij}, \boldsymbol{\delta}_{ij})$, $LM(\boldsymbol{\gamma}_{ij})$ and $LM(\boldsymbol{\delta}_{ij})$, results in the statistics which can also be applied in the framework of multiple populations and incomplete designs. For the tests for the shape of the ICC's, $LM(\boldsymbol{\gamma}_i, \boldsymbol{\delta}_i)$, $LM(\boldsymbol{\gamma}_i)$ and $LM(\boldsymbol{\delta}_i)$, some additional provisions need to be made. This has to do with the fact that the definition of the alternative model depends on the respondents sum scores $r(\boldsymbol{x}^{(i)}_n)$, which, in turn, depend on the partial response patterns $\boldsymbol{x}^{(i)}_n$. However, when every person responds to a unique set of items, setting

boundary scores for partitioning the score continuum becomes quite difficult, because the rationale of the procedure is that respondents grouped together should be approximately located in the same region of the latent ability space. Choosing boundary scores related to the proportion of correct responses is very crude, because the proportion of correct scores not only depends on the persons' ability, but also on the difficulty of the items. Partitioning the latent ability continuum and then deriving boundary scores for the respondents is extremely laborious and completely undermines the philosophy of the approach. Therefore, application of these statistics must be confined to designs were the sample of respondents is split up into a number of groups of respondents who were administered the same set of items. The sets of items are often called booklets. Then, for every booklet, boundary scores are set in such a way that resulting score ranges roughly reflect comparable ability levels across booklets. This approach is the same as the approach of the comparable S_i-test for the Rasch model (Glas & Verhelst, 1995).

A Numerical Example

The aim of this section is to give an example of the use of LM tests and modification indices using real data. The data are a completely random sample of the data emanating from the central national examinations in secondary education in the Netherlands in 1995. The items used are from a test concerning reading comprehension in English. To keep the presentation compact, only the first 10 items of this examination will be used. However, the results did prove typical for the complete examination. In Table 4, an overview of the data and the MML estimates are given. The second and third column, labeled "p-value" and "rit" contain the observed proportion correct scores for the items and the item-test correlations. The frequency distribution of the respondents unweighted sum scores is displayed in the last column. The remaining columns contain MML estimates of the parameters and estimates of their standard errors. The columns labeled "$Se^*(.)$" contain standard errors computed using the observed information matrix, given by (11), the columns labeled "$Se(.)$" contain standard errors computed using the Fischer information matrix, given by (13). It can be seen that these two estimates of the standard errors are very close indeed.

Insert Table 4 about here

Next, for these 10 items, LM statistics were computed, an overview is given in Table 5. In this example, for every item, the score range was divided into four sections. There are several considerations pertaining to the choice of the number of subsets S_i of the score range

and the choice of the boundary scores. Generally speaking, the number of score groups will depend on the number of items and the number of respondents available. Inspection of (22) and (24) reveals that, for polytomous items, the first order derivatives $\boldsymbol{h}(\boldsymbol{\delta}_i)$ are the difference between the number of persons obtaining an unweighted sum score in category s and scoring in category g of item i, $\sum_{n|w(s,\mathbf{x}_n^{(i)})=1} x_{nig}$ and its expected value.

Insert Table 5 about here

For dichotomous items, this boils down to the difference between the number of persons in s making the item correct and its expected value. Therefore, it may be a good strategy to form the subgroups in such a way that the observed and expected frequencies are not to low, which can be supported by setting the boundary scores in such a way that the numbers of respondents in each subgroup are comparable. As a side line, it must be mentioned that the fact that the magnitude of $\boldsymbol{h}(\boldsymbol{\delta}_i)$ depends on a difference between observed and expected frequencies will be helpful in assessing the severity of the model violations. Due to a large sample size, $\boldsymbol{\delta}_i$ may differ significantly from zero, yet the severity of the violation in terms of a difference between the observed and expected frequencies may be insignificant from a practical point of view.

The columns labeled "$LM^*(.)$" contain values for LM statistics computed using exact expressions for the matrix of second order derivatives. Since four subgroups were formed, $LM^*(\gamma_i,\delta_i)$ has 6 degrees of freedom and $LM^*(\gamma_i)$ and $LM^*(\delta_i)$ both have 3 degrees of freedom. To keep the presentation concise, association between items was evaluated for consecutive items only, the results are displayed in the second panel of Table 5. Here, $LM^*(\gamma_{ij},\delta_{ij})$ has 2 degrees of freedom and $LM^*(\gamma_{ij})$ and $LM^*(\delta_{ij})$ both have one degree of freedom. Finally, the columns of Table 5 labeled "$LM(\delta_{ij})$" and "$LM(\delta_i)$" contain LM statistics computed using the Fischer information matrix. Inspection shows that the values of these statistics are very close to the values obtained using the exact expressions for the second order derivatives. This result was typical for all analyses made in this numerical example. Therefore, in this section no further comparisons between the two approaches will be presented.

Although the primary aim of the tests presented here is to serve as item-oriented diagnostic tools, they also serve the purpose of evaluation of global model fit, especially if the number of items is large. Consider the ten significance probabilities of the $LM(\gamma_i,\boldsymbol{\delta}_i)$ test displayed in the third column of Table 5. Under the null-model, that is, under the 2-PL model, these ten significance probabilities should have an approximate uniform distribution.

Of course, this is only an approximation, because these 10 statistics are dependent. For reasons of dependence, one should not combine the significance probabilities of the $LM^*(\gamma_{ij}, \delta_{ij})$- , the $LM^*(\gamma_{ij})$- and $LM^*(\delta_{ij})$-statistics, because the dependence between these three statistics is too prominent. This can, for instance, be verified by inspection of Table 5. The same line of argument also applies to the three tests focussed at dependence between the items. Although the requirement of independence is also not fulfilled within one LM test replicated over items, here the dependence is far less prominent, and a fair approximation to the uniform distribution favors the model. On the other hand, a majority of low significance probabilities is indicative for global model violation. If, for instance, a significance level of 10% is used, a percentage of significant tests that greatly exceeds 10% is an indication of poor global model fit.

Insert Table 6 about here

The next interesting question is whether the results of Table 5 are much different in a Bayes model framework. Two analyses were made, one analysis where the discrimination parameters are assumed to be drawn from a known log-normal distribution, and an empirical Bayes analysis where the conjugate prior is introduced to this log-normal distribution. In the first case, the parameters of the log-normal distribution are fixed to $\mu_{\ln\gamma} = .0$ and $\sigma_{\ln\gamma} = 0.5$. The reason here is that these are the default values in Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, R.D., 1996), which will probably be the software mostly used by practitioners. Above, it was already mentioned that the conjugate prior for the normal distribution is normal for $\mu_{\ln\gamma}$ given $\sigma_{\ln\gamma}$ and inverted Wishart for $\sigma_{\ln\gamma}$ (Ando & Kaufman, 1965) Using the terminology of Mislevy (1986), for this last distribution the parameters $m = 5$ and $b = 1$ were chosen. The results of computation of the LM statistics are shown in Table 6. Generally, the pattern of significant indices remains the same. For instance, using a 10% significance level, for all three analyses, item 7 has a significant $LM(\gamma_i, \delta_i)$ and $LM(\delta_i)$. In the same manner, the item pair 3 and 4 has a significant $LM(\gamma_{ij}, \delta_{ij})$ and $LM(\gamma_{ij})$ and $LM(\delta_{ij})$-test in all three analyses. However, sometimes the pattern changes. For instance, the significant $LM(\gamma_i, \delta_i)$-test for item 3 disappears in the Bayesian analyses and a significant $LM(\gamma_{ij}, \delta_{ij})$-test for the item pair 2 and 3 appears in the empirical Bayes analysis. It must be mentioned that such changes occurred less when the number of items in the analysis was higher.

Insert Figure 1 about here

Insert Table 7 about here

In Section 3, it was sketched that using (16) an estimate of a freed fixed parameter can be computed by performing one Newton-Raphson step. Standard errors of these one-step estimates can be computed using the diagonal elements of $\boldsymbol{W}$. In Section 4, the alternative model was identified by imposing $\gamma_{igS_i} = \delta_{igS_i} = 0$. Therefore, for $s = 1, ..., S_i - 1$, the parameters γ_{igs} and δ_{igs} can be viewed as the deviations from the discrimination and difficulty parameter of group S_i, respectively. However, in practice it proves more elegant to have confidence intervals for all S_i score levels. Therefore, the MML estimates of α_{ig} and β_{ig} will be imputed in the alternative model as a fixed constants, so that the parameters γ_{igs} and δ_{igs} can be viewed as the deviations from these estimates for all groups $s = 1, ..., S_i$. This alternative parametrization entails that, for the computation of LM tests and modification indices, the elements of $\boldsymbol{h}(\boldsymbol{\xi})$ and $\boldsymbol{H}(\boldsymbol{\xi}, \boldsymbol{\xi})$ associated with α_{ig} and β_{ig} should be removed. Because this is just a simple reparametrization of the alternative model, this operation does not alter the outcome of the LM test.

In Table 7, one-step estimates are computed for the first two items, the results are displayed under the heading "Modification Indices" in Table 7. Assuming asymptotic normality of these estimates, they can be transformed into standardized normal indices. An example using the two items of Table 7 is shown in Figure 1. The circles signify standardized one-step estimates of γ_i, the triangles the standardized one-step estimates of δ_i. Using these displays, the locus of miss-fit can be identified at a glace. For instance, the lack of fit of the second item is mainly due to the low score level. Of course, an interesting question is how much the freed parameters will change if new MML estimates are computed, both for the parameters of the initial 2-PL model and the parameters of interest. In Table 7, these estimates are displayed under the heading "Parameter Estimates". It can be seen that these estimates are little different from the estimates under the heading "Modification Indices". The parameters of item 2 in group 4 seem an exception, but this appearant effect vanishes when the estimates are standardized by their standard error. The fact that new MML estimates were computed for all parameters in the model supports computing a likelihood ratio test. The log-likelihood of the original model equaled -12028.783, the model with additional parameters for item 1 resulted in -12027.109, the model with additional parameters for item 2 resulted in -12023.459. So, the LR-test for the first item has a value 3.338 (df=6, p=0.764), the test for the second item has a value

10.648 (df=6, p=0.100). This is in accordance with the other results, the first item seems to fit, the second might be called a borderline case. The strategy used here may serve as a prototype: first, compute LM tests and modification indices, which can be done quickly without additional estimation, then perform additional estimation for items where model fit appears to be troublesome, and, finally, relax the model in cases of serious violations.

A Power Study

In this section, an unassuming power study will be presented. It will in no way be exhaustive, because that would need a systematic variation of sample size, test length, parameter values and model violations which is far beyond the scope of this paper. The main purpose of the study was to get a general idea of the power of the LM tests. The arrangement of the study reported is quite arbitrary. However, the results are not significantly different from some other simulation studies that ware carried out. The sample size was equal to 1000 respondents, 9 dichotomous items were used. The discrimination parameters γ_i were equal to (.5,.5,.5,1,1,1,1.5,1.5,1.5). The difficulties δ_i were (-1,0,1,-1,0,1,-1,0,1). The ability distribution was standard normal. The first collection of studies was focussed on the tests for the shape of the ICC's, the second collection of studies was focussed on the tests for local independence.

Insert Table 8 about here

The results of the first collection of studies are reported in Table 8. In these studies, the ICC of item 5 was contaminated by introducing parameters γ_{is} and δ_{is}, $s = 1, ..., 4$. First, values were set for some parameter γ_{i*} and δ_{i*}, these values are shown in the third and fourth column of Table 8 under the labels "γ_{i*}" and "δ_{i*}". Using these values, two patterns of violations were created, $(\gamma_{i1}, \gamma_{i2}, \gamma_{i3}, \gamma_{i4}, \delta_{i1}, \delta_{i2}, \delta_{i3}, \delta_{i4})$ was equal to $(-\gamma_{i*}, \gamma_{i*}, \gamma_{i*}, -\gamma_{i*}, -\delta_{i*}, \delta_{i*}, -\delta_{i*}, \delta_{i*})$ in the first version, and equal to $(\gamma_{i*}, -\gamma_{i*}, -\gamma_{i*}, \gamma_{i*}, \delta_{i*}, -\delta_{i*}, \delta_{i*}, -\delta_{i*})$ in the second version. For an example, consider Table 8, where every row corresponds to a simulation study. Consider study 18. In the second column it can be seen that this study has the second pattern of violations, the third and fourth column display that $\gamma_{i*} = 0.50$ and $\delta_{i*} = 0.50$, so here $(\gamma_{i1}, \gamma_{i2}, \gamma_{i3}, \gamma_{i4}, \delta_{i1}, \delta_{i2}, \delta_{i3}, \delta_{i4})$ was equal to $(-0.50, 0.50, 0.50, -0.50, -0.50, 0.50, -0.50, 0.50)$. Using this setup, 100 replications were made for every row in Table 8 , for every replication 1000 response patterns were generated, MML estimates were computed and $LM(\gamma_i, \delta_i)$-, $LM(\gamma_i)$- , and $LM(\delta_i)$-tests were performed using a 10% significance level. The proportion of significant

tests is displayed in the last six columns of Table 8, the tests in the columns labeled $LM^*(.)$ were computed using the observed information matrix, the tests in the columns labeled $LM(.)$ were computed using the Fischer information matrix. The first row of Table 8 corresponds to the null-model, that is, to the 2-PL model, and it can be seen that the proportion of tests significant at 10% is approximately equal to 0.10, which is as it should be. Further, it can be seen that the proportions of significant tests are monotone increasing in γ_{i*} and δ_{i*}, which is also in accordance with the purpose of the tests. However, an interesting feature of the results is that all tests are sensitive to all violations, for instance, $LM(\gamma_i)$ is both sensitive to a violation $\gamma_{i*} = 0$ and $\delta_{i*} \neq 0$, and a violation $\gamma_{i*} \neq 0$ and $\delta_{i*} = 0$. In fact, the power of $LM(\delta_i)$ to a violation $\gamma_{i*} \neq 0$ and $\delta_{i*} = 0$ is greater than the power of $LM(\gamma_i)$. So it must be concluded that attribution of the outcome of the test to specific parameters will be quite difficult. This result must be attributed to the high correlation between estimates of the item discrimination and item difficulty parameters, so the reason for the poor discriminative power of the tests must be attributed to the properties of the 2- PL model, and not to the properties of the tests. Summing up, the tests must be used as caution indices, and one must not expect to be able to trace significant results back either the item discrimination or difficulty parameters.

Insert Table 9 about here

Table 9 contains the results of a comparable study to the power of $LM(\gamma_{ij},\delta_{ij})$, $LM(\gamma_{ij})$ and $LM(\delta_{ij})$. Again, the number of items is 9 and the number of respondents is 1000. Also the parameters of the 2- PL model were the same as in the previous study. In Table 9, every row of the table corresponds to a study. Association between items was induced by introducing additional parameters γ_{ij} and δ_{ij}, in the second and third column it can be seen that one half of the studies the concerns association between item 1 and 5, the other half concerns association between item 5 and 8. The values of γ_{ij} and δ_{ij} are displayed in the next two columns. For every study, 100 replications were made and the proportion of LM tests significant at the 10% level was computed. The results are given in the last 6 rows of Table 9. As above, the tests in the columns labeled $LM^*(.)$ were computed using the observed information matrix, the tests in the columns labeled $LM(.)$ were computed using the Fischer information matrix. Contrary to the above studies, the tests prove more discriminative with respect to the specific violation imposed. So $LM(\gamma_{ij})$ has substantial power for a violation $\gamma_{ij} = 0$ and $\delta_{ij} \neq 0$, , while the power of $LM(\gamma_{ij})$ is low. Analogously, for a violation $\gamma_{ij} \neq 0$ and $\delta_{ij} = 0$ the opposite applies.

Discussion

In the present article, it was shown that LM tests and modification indices are a practical and useful tool for evaluation of model fit. Their practicality is a result of the circumstance that most of the ingredients needed are available at the end of the estimation procedure, so very little additional computations have to be made. They are usefulness because they are item oriented diagnostic tools, which give an indication of the source of model violations. Potentially, they offer the possibility of directed model relaxation to obtain sufficient model fit. On the other hand, the discriminative power of the approach must not be exaggerated. For instance, if the model is grossly violated, a sum score $r(\boldsymbol{x}_n^{(i)})$ on a partial response vector $\boldsymbol{x}_n^{(i)}$ may no longer be a valid indication of ability, so that the underpinning of the $LM(\gamma_i)$-, $LM(\delta_i)$- and $LM(\gamma_i, \delta_i)$-test becomes unrealistic. Further, the discriminative power of the tests is, of course, also limited by the characteristics of the model, for instance, the power study made apparent that the well-known dependence between α_{ig}, $g = 1, ..., m_i$ and β_{ig}, $g = 1, ..., m_i$ obstructed the attribution of model violations to either set of parameters. An advantageous aspect of some of the statistics is that they are based on a difference between observed and expected frequencies, so the importance of a significant model test can be assessed in a framework that is directly related to observed data. The approach presented here can obviously be extended in several directions. The first extension is tailoring the approach to the 3-PL model. Further, the model can also be extended to encompass models with multidimensional ability distributions. Finally, in many structural models on ability parameters, the item parameters estimates issued from a calibration phase are imputed into the structural model as known constants. Also evaluation of the validity of these imputed constants when confronted with the new data seems another promising area where LM statistics and modification indices might be useful.

References

Aitchison,J. & Silvey, S.D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics 29,* 813-828

Andersen, E.B. (1973). A goodness of for test for the Rasch model. *Psychometrika, 38,* 123-140.

Ando, A. & Kaufmann, O.M. (1965). Bayesian analysis of the independent normal process-neither mean nor precision known. *Journal of the American Statistical Association, 60,* 347-358.

Birnbaum, A. (1968). Some latent trait models. (hoofdstuk 17 in:) F.M. Lord & M.R. Novick, *Statistical theories of mental test scores.* Addison-Wesley: Reading (Mass.).

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

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). *Einf?hrung in die Theorie Psychologischer Tests* [Introduction to the theory of psychological tests.] Bern: Huber.

Follmann, D. (1988). Consistent estimation in the Rasch model based on non-parametric margins. *Psychometrika, 53,* 553-562.

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. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson, (Ed.), *Objective measurement: theory into practice, Vol. 1,* (pp.236-258), New Jersey: Ablex Publishing Corporation.

Glas, C.A.W. (1997). Testing the generalized partial credit model. In M. Wilson, G. Engelhard, Jr., and K. Draney (Eds.), *Objective measurement: theory into practice, Vol. 4,* New Jersey: Ablex Publishing Corporation.

Glas, C.A.W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica,* to appear.

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

Glas, C.A.W., & Verhelst, N.D. (1995). Tests of fit for polytomous Rasch models. In G.H.Fischer & I.W.Molenaar (eds.). *Rasch models. Their foundation, recent developments and applications.* New York: Springer.

Hemker, B.T., Sijtsma, K, Molenaar, I.W. & Junker, B.W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61,* 679-693.

Jannarone, R.J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51,* 357-373.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49,* 223-245.

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

DeLeeuw, J., & Verhelst, N. D. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11,* 183-196.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44,* 226-233.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, N.J., Erlbaum.

Martin-Löf, P. (1973). *Statistika Modeller. Anteckningar från seminarier Lasåret 1969-1970, utardeltade av Rolf Sunberg. Obetydligt ändrat nytryck, oktober 1973.* Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.

Martin Löf, P. (1974). The notion of redundancy and its use as a quantitative measure if the discrepancy between a statistical hypothesis and a set of observational data. *Scandinavian Journal of Statistics, 1,* 3-18.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Mislevy, R.J. & Bock, R.D. (1990). *PC-Bilog. Item analysis and test scoring with binary logistic models.* Chicago: Scientific Software International, Inc.

Molenaar, I.W. (1983). Some improved diagnostics for failure in the Rasch model. *Psychometrika, 48,* 49-72.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Rao, C.R. (1947). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50-57.

Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika, 61,* 509-528.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Thissen, D. (1991). *MULTILOG. Multiple, categorical item analysis and test scoring*

using item response theory. Chicago: Scientific Software International, Inc.

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

Zimowski, M.F., Muraki, E., Mislevy, R.J. & Bock, R.D. (1996). *Bilog MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items.* Chicago: Scientific Software International, Inc.

Table 1
Expressions for $B_n(\phi_a, \phi_b)$ for the parameters of item i

	γ_{ihs} or α_{ih}	γ_{igs} or α_{ig}	δ_{ihs} or β_{ih}	δ_{igs} or β_{ig}
γ_{ihs}	$-\theta_n^2\psi_{nihs}(1-\psi_{nihs})$	$\theta_n^2\psi_{nihs}\psi_{nigs}$	$\theta_n\psi_{nihs}(1-\psi_{nihs})$	$-\theta_n\psi_{nihs}\psi_{nigs}$
γ_{igs}	$\theta_n^2\psi_{nigs}\psi_{nihs}$	$-\theta_n^2\psi_{nigs}(1-\psi_{nigs})$	$-\theta_n\psi_{nigs}\psi_{nihs}$	$\theta_n\psi_{nigs}(1-\psi_{nigs})$
δ_{ihs}	$\theta_n\psi_{nihs}(1-\psi_{nihs})$	$-\theta_n\psi_{nihs}\psi_{nigs}$	$-\psi_{nihs}(1-\psi_{nihs})$	$\psi_{nihs}\psi_{nigs}$
δ_{igs}	$-\theta_n\psi_{nigs}\psi_{nihs}$	$\theta_n\psi_{nigs}(1-\psi_{nigs})$	$\psi_{nigs}\psi_{nihs}$	$-\psi_{nigs}(1-\psi_{nigs})$

Table 2
Cross-tabulation of Probabilities

$Pr(x_i, x_j \mid \alpha_i, \alpha_j, \beta_i, \beta_j, \gamma_{ij}, \delta_{ij}) \propto$

	$x_i = 0$	$x_i = 1$
$x_j = 0$	1	$\exp(\alpha_i\theta - \beta_i)$
$x_j = 1$	$\exp(\alpha_j\theta - \beta_j)$	$\exp((\alpha_i + \alpha_j + \gamma_{ij})\theta - \beta_i - \beta_i + \delta_{ij})$

Table 3
Expressions for $B_n(\phi_a, \phi_b)$ for the parameters of item i

	γ_{ih}	γ_{ig}	δ_{ih}	δ_{ig}
γ_{igjh}	$-\theta_n^2\zeta_{nigjh}(1-\zeta_{nigjh})$	$\theta_n^2\zeta_{nigjh}\zeta_{nikjl}$	$\theta_n\zeta_{nigjh}(1-\zeta_{nigjh})$	$-\theta_n\zeta_{nigjh}\zeta_{nikjl}$
γ_{ikjl}	$\theta_n^2\zeta_{nikjl}\zeta_{nigjh}$	$-\theta_n^2\zeta_{nikjl}(1-\zeta_{nikjl})$	$-\theta_n\zeta_{nikjl}\zeta_{nigjh}$	$\theta_n\zeta_{nikjl}(1-\zeta_{nikjl})$
δ_{igjh}	$\theta_n\zeta_{nigjh}(1-\zeta_{nigjh})$	$-\theta_n\zeta_{nigjh}\zeta_{nikjl}$	$-\zeta_{nigjh}(1-\zeta_{nigjh})$	$\zeta_{nigjh}\zeta_{nikjl}$
δ_{ikjl}	$-\theta_n\zeta_{nikjl}\zeta_{nigjh}$	$\theta_n\zeta_{nikjl}(1-\zeta_{nikjl})$	$\zeta_{nikjl}\zeta_{nigjh}$	$-\zeta_{nikjl}(1-\zeta_{nikjl})$
α_{ig}	$-\theta_n^2\zeta_{nigjh}(1-\sum_l\zeta_{nigjl})$	$-\theta_n^2\zeta_{nikjl}(1-\sum_h\zeta_{nikjl})$	$\theta_n\zeta_{nigjh}(1-\sum_l\zeta_{nigjl})$	$\theta_n\zeta_{nigjh}(1-\sum_h\zeta_{nigjh})$
α_{ik}	$\theta_n^2\zeta_{nigjh}\sum_l\zeta_{nikjl}$	$\theta_n^2\zeta_{nikjl}\sum_h\zeta_{nikjh})$	$-\theta_n\zeta_{nigjh}\sum_l\zeta_{nikjl}$	$-\theta_n\zeta_{nikjl}\sum_h\zeta_{nikjh})$
β_{ig}	$\theta_n\zeta_{nigjh}(1-\sum_l\zeta_{nigjl})$	$\theta_n\zeta_{nikjl}(1-\sum_h\zeta_{nigjh})$	$-\zeta_{nigjh}(1-\sum_l nigjl)$	$-\zeta_{nikjl}(1-\sum_h\zeta_{nigjh})$
β_{ik}	$-\theta_n\zeta_{nigjh}\sum_l\zeta_{nikjl}$	$-\theta_n\zeta_{nikjl}\sum_h\zeta_{nikjh}$	$\zeta_{nigjh}\sum_l\zeta_{nikjl}$	$\zeta_{nikjl}\sum_h\zeta_{nikjh}$

Table 4
Data Summary and MML parameter estimation of 10 Examination Items
Number of observations = 2039

item	p-value	rit	α_i	β_i	$Se^*(\alpha_i)$	$Se^*(\beta_i)$	$Se(\alpha_i)$	$Se(\beta_i)$	score	frequency
									0	2
1	.40	.36	.30	.40	.071	.047	.071	.047	1	7
2	.86	.41	1.28	-2.31	.176	.145	.169	.138	2	23
3	.87	.37	.95	-2.16	.132	.105	.132	.105	3	92
4	.49	.41	.50	.06	.075	.047	.077	.047	4	175
5	.81	.39	.75	-1.59	.103	.074	.106	.075	5	314
6	.57	.42	.59	-.32	.078	.049	.081	.049	6	380
7	.66	.39	.53	-.71	.080	.051	.082	.051	7	425
8	.63	.47	.85	-.61	.097	.055	.100	.056	8	333
9	.62	.40	.49	-.52	.078	.049	.079	.049	9	224
10	.56	.43	.63	-.25	.083	.049	.083	.049	10	64

Table 5
LM modification indices for 10 Examination Items

	item	$LM^*(\gamma_i,\delta_i)$	p	$LM^*(\gamma_i)$	p	$LM^*(\delta_i)$	p	$LM(\delta_i)$	p
	1	3.26	.78	.87	.83	1.36	.72	1.41	.70
	2	11.93	.06	12.64	.01	2 48	.48	2.58	.46
	3	13.29	.04	5.48	.14	.75	.86	1.41	.70
	4	2.88	.82	1.12	.77	.45	.93	.46	.93
	5	4.43	.62	2.29	.52	1.02	.80	.90	.82
	6	7.47	.28	2.47	.48	4.29	.23	5.00	.17
	7	11.62	.07	4.30	.23	9.20	.03	9.70	.02
	8	7.31	.29	3.63	.30	1.53	.67	1.52	.68
	9	11.10	.09	5.51	.14	6.10	.11	6.30	.10
	10	9.15	.17	6.56	.09	3.76	.29	4.23	.24
item i	item j	$LM^*(\gamma_{ij},\delta_{ij})$	p	$LM^*(\gamma_{ij})$	p	$LM^*(\delta_{ij})$	p	$LM(\delta_{ij})$	p
1	2	4.25	.12	3.39	.07	.62	.43	.64	.42
2	3	.46	.80	.44	.51	.41	.52	.41	.52
3	4	18.91	.00	4.93	.03	19.73	.00	18.69	.00
4	5	.85	.65	.80	.37	.31	.58	.31	.58
5	6	1.89	.39	1.74	.19	.18	.67	.18	.67
6	7	.89	.64	.35	.55	.27	.61	.26	.61
7	8	3.85	.15	3.59	.06	.16	.69	.16	.69
8	9	4.64	.10	.91	.34	2.40	.12	2.22	.14
9	10	2.41	.30	1.73	.19	.24	.62	.23	.63

Table 6
LM modification indices in a Bayesian Framework

Statistics Computed Using Fixed Prior							
	item	$LM(\gamma_i,\delta_i)$	p	$LM(\gamma_i)$	p	$LM(\delta_i)$	p
	1	4.99	.55	2.45	.48	1.97	.58
	2	11.83	.07	6.78	.08	2.56	.47
	3	4.94	.55	3.58	.31	1.28	.73
	4	1.63	.95	.64	.89	.64	.89
	5	3.31	.77	2.23	.53	1.14	.77
	6	8.37	.21	2.66	.45	5.12	.16
	7	11.88	.06	5.54	.14	10.82	.01
	8	5.30	.51	2.60	.46	1.58	.66
	9	11.73	.07	6.09	.11	7.21	.07
	10	9.91	.13	6.25	.10	4.70	.20
item i	item j	$LM(\gamma_{ij},\delta_{ij})$	p	$LM(\gamma_{ij})$	p	$LM(\delta_{ij})$	p
1	2	1.59	.45	.20	.66	1.57	.21
2	3	.30	.86	.27	.60	.11	.74
3	4	18.49	.00	7.61	.01	17.78	.00
4	5	2.06	.36	2.03	.15	.13	.72
5	6	.58	.75	.53	.46	.27	.61
6	7	.09	.96	.01	.92	.08	.77
7	8	1.85	.40	1.79	.18	.38	.54
8	9	2.04	.36	.18	.67	1.50	.22
9	10	.42	.81	.31	.58	.06	.81
Statistics Computed Using Emperical Prior							
	item	$LM(\gamma_i,\delta_i)$	p	$LM(\gamma_i)$	p	$LM(\delta_i)$	p
	1	15.22	.02	4.55	.21	2.24	.52
	2	8.47	.21	7.26	.06	3.90	.27
	3	8.25	.22	4.86	.18	1.49	.68
	4	1.21	.98	.41	.94	.67	.88
	5	4.08	.67	1.15	.76	1.26	.74
	6	7.68	.26	2.32	.51	5.27	.15
	7	11.59	.07	4.50	.21	10.67	.01
	8	6.79	.34	1.31	.73	1.37	.71
	9	11.31	.08	4.89	.18	7.52	.06
	10	9.50	.15	4.68	.20	4.80	.19
item i	item j	$LM(\gamma_{ij},\delta_{ij})$	p	$LM(\gamma_{ij})$	p	$LM(\delta_{ij})$	p
1	2	1.33	.51	.01	.93	1.01	.32
2	3	11.37	.00	.48	.49	.94	.33
3	4	17.20	.00	5.75	.02	16.81	.00
4	5	1.55	.46	1.49	.22	.04	.84
5	6	.98	.61	.91	.34	.44	.51
6	7	.09	.96	.00	.98	.08	.78
7	8	2.65	.27	2.62	.11	.32	.57
8	9	2.52	.28	.55	.46	1.44	.23
9	10	.70	.70	.58	.45	.05	.83

Table 7
Modification Indices and Parameter Estimation

		Modification Indices				Parameter Estimates			
item	group	γ_{is}	δ_{is}	$se(\gamma_{is})$	$se(\delta_{is})$	γ_{is}	δ_{is}	$se(\gamma_{is})$	$se(\delta_{is})$
1	1	-.169	.129	.197	.152	-.159	.124	.193	.153
	2	.013	-.039	.669	.099	.013	-.039	.665	.099
	3	.230	.190	.894	.357	.243	.198	.955	.407
	4	-.413	-.432	.534	.468	-.408	-.424	.516	.447
2	1	-.625	.642	.301	.290	-.557	.555	.225	.198
	2	.412	-.405	1.865	.721	.588	-.510	2.148	1.110
	3	-.272	.187	1.889	.254	-.274	.182	1.387	.211
	4	1.909	-.345	3.432	.870	.391	-.542	3.860	.399

Table 8
Study of the Power of the Test for the Shape of the ICC
100 Replications per Study

study	pattern	γ_{i*}	δ_{i*}	$LM^*(\gamma_i,\delta_i)$	$LM(\gamma_i,\delta_i)$	$LM^*(\gamma_i)$	$LM(\gamma_i)$	$LM^*(\delta_i)$	$LM(\delta_i)$
0	0	.00	.00	.10	.09	.11	.10	.10	.10
1	1	.10	.00	.16	.06	.13	.13	.11	.10
2		.25	.00	.29	.36	.17	.23	.32	.33
3		.50	.00	.74	.77	.50	.56	.80	.81
4	2	.10	.00	.29	.13	.17	.17	.13	.09
5		.25	.00	.30	.19	.25	.17	.30	.26
6		.50	.00	.76	.72	.63	.52	.84	.76
7	1	.00	.10	.23	.22	.25	.21	.25	.24
8		.00	.25	.78	.78	.52	.58	.76	.81
9		.00	.50	1.00	1.00	1.00	1.00	1.00	1.00
10	2	.00	.10	.31	.23	.25	.23	.32	.26
11		.00	.25	.85	.82	.78	.75	.88	.86
12		.00	.50	1.00	1.00	.97	.97	1.00	1.00
13	1	.10	.10	.32	.39	.27	.28	.34	.37
14		.25	.25	.97	.97	.84	.89	.99	.99
15		.50	.50	1.00	1.00	.99	.99	1.00	1.00
16	2	.10	.10	.40	.37	.34	.30	.47	.44
17		.25	.25	.95	.96	.93	.90	1.00	.99
18		.50	.50	1.00	1.00	1.00	1.00	1.00	1.00

Table 9
Study of the Power of the Test for Association between items
100 Replications per Study

study	item i	item j	γ_{ij}	δ_{ij}	$LM^*(\gamma_{ij},\delta_{ij})$	$LM(\gamma_{ij},\delta_{ij})$	$LM^*(\gamma_{ij})$	$LM(\gamma_{ij})$	$LM^*(\delta_{ij})$	$LM(\delta_{ij})$
0	0	0	.00	.00	.09	.09	.08	.08	.11	.12
1	1	5	.05	.00	.10	.10	.10	.10	.13	.14
2			.10	.00	.13	.14	.13	.13	.14	.14
3			.25	.00	.24	.22	.22	.21	.10	.10
4			.50	.00	.55	.57	.68	.69	.11	.11
5	5	8	.05	.00	.15	.16	.13	.12	.07	.08
6			.10	.00	.12	.15	.17	.15	.12	.12
7			.25	.00	.21	.18	.19	.18	.13	.13
8			.50	.00	.36	.38	.46	.47	.15	.15
9	1	5	.00	.05	.13	.15	.06	.13	.14	.15
10			.00	.10	.13	.15	.10	.11	.17	.16
11			.00	.25	.35	.39	.13	.12	.49	.49
12			.00	.50	.92	.92	.13	.13	.96	.96
13	5	8	.00	.05	.13	.16	.06	.06	.17	.19
14			.00	.10	.17	.21	.14	.14	.22	.21
15			.00	.25	.35	.38	.09	.10	.38	.42
16			.00	.50	.90	.91	.14	.15	.93	.94
13	1	5	.05	.05	.11	.11	.11	.12	.09	.10
13			.10	.10	.12	.13	.13	.14	.15	.17
14			.25	.25	.57	.58	.40	.39	.59	.60
15			.50	.50	.97	.97	.90	.90	.93	.93
16	5	8	.05	.05	.11	.11	.07	.07	.11	.13
16			.10	.10	.17	.17	.10	.11	.15	.15
17			.25	.25	.41	.41	.17	.18	.44	.47
18			.50	.50	.87	.87	.51	.54	.83	.84

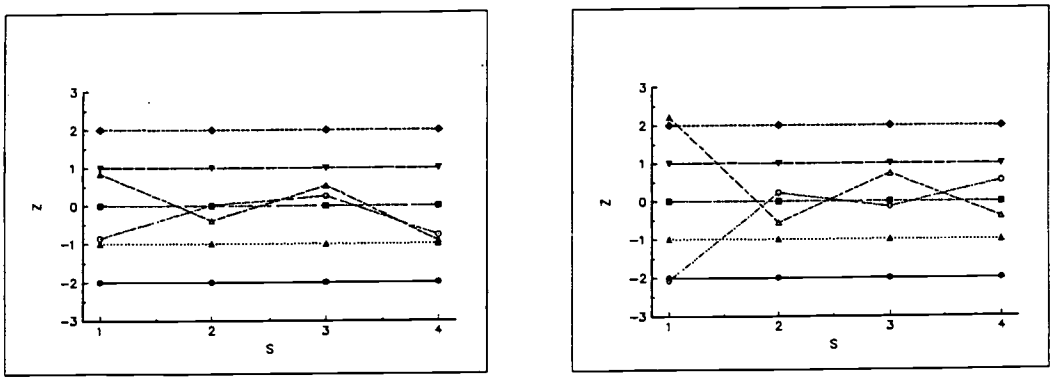


Figure 1. Graphic Display of the Efficient Score Test for Two Items.

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-98-04	C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*
RR-98-03	C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*
RR-98-02	R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*
RR-98-01	C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*
RR-97-07	H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*
RR-97-06	H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*
RR-97-05	W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*
RR-97-04	W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*
RR-97-03	W.J. van der Linden, *Multidimensional Adaptive Yesting with a Minimum Error-Variance Criterion*
RR-97-02	W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*
RR-97-01	W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*
RR-96-04	C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*
RR-96-03	C.A.W. Glas, *Testing the Generalized Partial Credit Model*
RR-96-02	C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*
RR-96-01	W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*
RR-95-03	W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*
RR-95-02	W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*
RR-95-01	W.J. van der Linden, *Some decision theory for course placement*
RR-94-17	H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*
RR-94-16	H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*
RR-94-15	H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

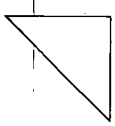
RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.



faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").